Információs Társadalom

TÁRSADALOMTUDOMÁNYI FOLYÓIRAT
Alapítva 2001-ben

Főszerkesztő és felelős szerkesztő: Héder Mihály

Lapterv: Szépkilátás Stúdió
Kiadványszerkesztés: VEGA^{2000} Bt.

Kiadja
Az INFONIA (Információs Társadalomért, Információs Kultúráért) Alapítvány és a Gondolat Kiadó

A folyóirat megjelenését támogatták:

A szerkesztőség levelezési címe
BME Filozófia és Tudománytörténet Tanszék
H-1111 Budapest, Egry József u. 1., E610
e-mail: inftars@filozofia.bme.hu

Felelős kiadó Bácskai István
Gondolat Kiadó
1088 Budapest, Szentkirályi u. 16. tel: 486-1527
www.gondolatkiado.hu

Készült a Rolling Site Nyomdában
ISSN 1587-8694

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

Lectori salutem!

Nagy örömünkre szolgál ez a pillanat: az olvasó az Információs Társadalom első angol kiadását tartja a kezében. A név marad magyarul, hiszen Information Society c. folyóirat már van a Taylor & Francis kiadásában. Az ISSN és egyéb azonosítók is változatlanok, a folytonosság jegyében. Külföldi szerzőinknek "InfTars" néven emlegetjük a folyóiratot, amit ők át is vettek, ám ez csak becenév vagy "handle", formálisan minden maradt a régiben.

E szám tartalma nagyrészt, de nem teljesen a BME GTK által rendezett BudPT19 Technikafilozófia konferencia előadásai alapján készült. A tíz tanulmány izgalmas, változatos témákat ölel fel.

Jól ismert, hogy a gépi tanulás, különösen a neurális háló alapú alkalmazások egészen elképesztő teljesítményre képesek, azonban ennek ára van a transzparencia terén. Paul Grünke (Karlsruhe Institute of Technology) tanulmányban a gépi tanulás "epistemic opacity" fajtáinak osztályozásával foglalkozik, a híres AlphaZero-t használva példaként.

A második tanulmányban Héder Mihály (BME) egy empirikus kutatás eredményeit ismerteti. Héder egy 365 napos időszakban minden nap vizsgálta a lájk-komment-kattintás feketepiac egyik fontos platformját, így pontos számokkal tud szolgálni az árakról, igényekről, fenyegető trendekről.

Karakas Alexandra (ELTE) a technológia tervezői és társadalmi kontroll lehetőségeit vizsgálja a "malfunction", azaz a hibás működés értelmezésein keresztül, ezzel áttételesen a funkció-vitákhoz járulva hozzá.

Aleksandra Kazakova (Gubkin Russian State University) az orvosbiológiai mérnökképzés etikai megközelítéseit vizsgálja, az Egyesült Államok és Oroszország egyes képzéseinek mintatanterveit és kurzusleírásait ütköztetve egymással.

Chang-Yun Ku (Academia Sinica, Tajvan) azzal a kényes kérdéssel foglalkozik, hogy milyen esetekben bírálhatja felül a mesterséges intelligencia az emberi döntést, vagy legalábbis milyen feszültségek keletkeznek az ember-gép kollaborációban episztemikus és adatvédelmi dimenziókban.

Pintér Dániel Gergő és Ihász Péter Lajos a gépi nyelvfeldolgozás (NLP) vállalati kommunikációs eszközként való felhasználási opcióit elemzik, saját kommunikációs keretrendszer segítségével.

Szabó Krisztina (BME) a "képernyő korában" vizsgálja az olvasás új szerepét. Újraértékeli az írástudás különféle kulcsfogalmait, adaptálva a megváltozott körülményekre és részletesen jellemzi a 21-ik századi digitális olvasás jellemzőit és sajátosságait.

Szántó Zoltán Oszkár és társai (Corvinus) tanulmányukban a "Social Futuring", magyarul Társadalmi Jövőképesség Indexet mutatják be. A Társadalmi jövőképesség egy interdiszciplináris fejlettségi-jóléti mutató, amely egyben potenciált is próbál jósolni.

Radu Uszkai (Bucharest University of Economic Studies) a szexrobotok etikai kérdéseit vizsgálja rawls-i keretrendszerre alapozva. Uszkai az igen nehéz témát egy másik nehéz témával, a mentális és fizikai fogyatékossággal élők szexuális jogaival ötvözi, érvelését végig körültekintően építve.

Az utolsó tanulányt ebben a számban Anda Zahiu (Research Center in Applied Ethics, University of Bucharest) jegyzi. Kutatási kérdése az, hogy az immerzív virtuális valóságnak milyen hatása lehet az önképre. Ehhez az elmefilozófiában jól ismert kiterjesztett elme koncepciót veszi alapul, kiegészítve azt a VR sajátosságaival.

Kellemes olvasást kívánunk!

A szerkesztőség

Chess, Artificial Intelligence, and Epistemic Opacity

Paul Grünke

Abstract
In 2017 AlphaZero, a neural network-based chess engine shook the chess world by convincingly beating Stockfish, the highest-rated chess engine. In this paper, I describe the technical differences between the two chess engines and based on that, I discuss the impact of the modeling choices on the respective epistemic opacities. I argue that the success of AlphaZero's approach with neural networks and reinforcement learning is counterbalanced by an increase in the epistemic opacity of the resulting model.

1 Introduction

Games have always been a welcome area for AI developers to test and develop their newest techniques. Chess is the most famous game in this context and the one that has been studied the most by computer scientists. The fascination for a machine playing chess dates back to the late 18th century when a chess automaton was exhibited throughout Europe. Many of the great minds of early computer science such as Charles Babbage, Alan Turing and John von Neumann devised their own approaches towards a program that would be able to play chess. This culminated in the victory of IBM's "Deep Blue" computer versus the (at the time) reigning world champion, Garry Kasparov, in 1997, which was a major event for the artificial intelligence community. It also started the domination of computer engines in the game of chess; which has today developed so far that a match between a computer and a human player would not be interesting anymore[1] and all top players as well as many amateur players are relying very heavily on computers in their preparation and training (Chessentials 2019). Today's chess engines are very sophisticated programs that include specifically designed search algorithms and evaluation functions, incorporate opening books and endgame databases.

In 2017, 20 years after the success of Deep Blue, a new kind of chess engine has been introduced. AlphaZero is a chess engine based on a neural network created by British AI company DeepMind. Its only inputs were the rules of chess and within a few hours of training via playing against itself it became the strongest chess engine in the world. In this paper, I describe this new approach towards chess engines, discuss its differences from other approaches and investigate the hypothesis that the success of this approach with neural networks and reinforcement learning is counterbalanced by an increase in epistemic opacity of the resulting

[1] Hikaru Nakamura, at the time rated the sixth best player in the world, played a match against a strong chess engine with getting additional material or moves in 2016. He drew three games and lost one (Chabris 2016).

model. An increase in epistemic opacity usually leads to a decrease in our ability to understand and control the resulting model as well as limit our options of learning from it.

In the following sections, I first sketch the development of chess engines (Section 2), and then describe and compare AlphaZero and the strongest traditional chess engine (Section 3). In Section 4, I introduce the concept of epistemic opacity, compare the most advanced classical chess engine with AlphaZero with respect to their epistemic opacities, and discuss the different kinds of opacities that are involved and then conclude (Section 5).

2 History of Chess Engines and AI

In 1770, Wolfgang von Kempelen presented to the Habsburg Archduchess Maria Theresa what would be known as The Turk. It was a machine, which consisted of a wooden man sitting at a cabinet with a chessboard on top of it. The machine was able to move the chess pieces on the board seemingly autonomously. This in itself would not have been very impressive, if it were not for the fact that The Turk was a very good chess player defeating most of its challengers, including famous personalities of the time such as Napoleon Bonaparte, Benjamin Franklin and Charles Babbage. Theories about how the machine works developed quickly, sparked by the same disbelief that led the British author Philip Thicknesse to write: “That an automaton can be made to move the Chessmen properly, as a pugnacious player, in consequence of the preceding move of a stranger, who undertakes to play against it, is utterly impossible” (Thicknesse 1784). As it turned out, he was right for the time being since a small-statured player hidden in the wooden cabinet operated The Turk. Even though the Turk was a hoax it played an important role since it fascinated people with the idea of an intelligent machine and raised questions about the possibilities of thinking machines which are still relevant today (Morton 2015, Standage 2003).

Thicknesse’s statement was proven wrong for the first time in the 1950s. Already in the late 1940s both Alan Turing and Claude Shannon (Shannon 1950) created algorithms that were able to play chess. Since no computers with the ability to compute the algorithm were available yet, Turing tried the program in 1951 by calculating everything manually. Also in 1951 Dietrich Prinz, a colleague of Alan Turing, implemented a program which was able to solve mate-in-2 problems. Finally, in 1957 Alex Bernstein, an IBM engineer, implemented a program on a computer for the first time, which was able to play a game of chess (Chessentials 2019).

The first chess engines were not very strong and could be beaten easily by strong amateur players. During the next decades, the quality of the chess engines increased continuously due to developments both in hardware and software and this culminated in the most important event in computer chess history: the IBM computer Deep Blue beating the reigning world champion Garry Kasparov in 1997 in a match of six games. After having won matches against predecessors of the program, Kasparov lost this match 3.5-2.5. Once again, the human-vs-machine setup in chess sparked fascination as well as fear of the developments of artificial intelligence. This time legitimately, as a machine had become strong enough to beat the arguably best chess player of all time (Krauthammer 1997). In the years after the Deep Blue match, a few more matches were played between world-class chess players and chess engines with mixed results, but these matches stopped after 2006, when Vladimir Kramnik, Kasparov's successor as world champion, also lost a six game match. Chess engines now mainly play against each other in Chess Computer Championships. The winner of many of the most recent Computer Chess Tournaments and one of the highest-rated Chess Engines is an open-source program called Stockfish. The general programming approach of Stockfish is similar to the one of Deep Blue, using "sophisticated search techniques, domain-specific adaptations, and handcrafted evaluation functions that have been refined by human experts over several decades" (Silver et al. 2018).

In 2017, the AI company DeepMind presented a new kind of chess engine: AlphaZero. Its approach is radically different from the former chess engines; the only input given to AlphaZero were the rules of chess. Using reinforcement learning on a specifically tailored neural network, AlphaZero was then trained by playing against itself for nine hours. The resulting program played a match against Stockfish over 1000 games that it won convincingly (winning 155 games and losing 6). The result in itself is already surprising. Even more surprising to the chess community was the style of the new engine, which is radically different from previous ones. AlphaZero appears to play in a risky attacking style but nearly never runs the risk of losing. Besides that, AlphaZero introduced a number of new motifs and strategies in all stages of the game, which have already been adopted by elite chess players. Therefore, there are more than enough reasons for chess players to delve into the depths of AlphaZero, analyse its games and try to understand the reasons for its success. For philosophers of science, this is an opportunity to gain a better understanding of the success of neural networks and machine learning techniques and to discuss the possible limits inherent to this method.

3 Comparing AlphaZero and Stockfish – a new kind of chess engine

In this section, I compare the techniques used for programming AlphaZero and Stockfish respectively and highlight relevant differences. This will lay the foundation to assess their respective epistemic opacities in the next section.

Chess engines are generally based on two core components. They have a way to evaluate positions and they have a search algorithm that determines which moves are available and in which order they should be considered. For both of these components Stockfish and AlphaZero use significantly different approaches.

3.1 Stockfish

Stockfish represents chess positions by using a vector that has chess-specific features as elements. These are handcrafted and include elements, which represent for example the pieces each player still has left, the safety of the kings or the pawn structure. The programmers chose those features in cooperation with strong chess players. Each of these features has been included because it has proven relevant in human experience of playing chess and successful in test runs with the engine. Each of these features has a specific weight assigned to it and the evaluation of the position results from a sum of all the features multiplied with their weights. This evaluation is then output in pawn-equivalents.

+2.00 for example means that the player with the white pieces has an advantage that is evaluated equivalently to having two extra pawns (Silver et al. 2018).

In order to figure out what the best move in a specific position is, Stockfish spans a tree of possible developments and evaluates the resulting positions using the evaluation function. This evaluation is only applied to "quiet" positions, i.e. positions without unresolved captures or checks. Therefore, once the desired depth of calculation is reached, there is a quiescence search[2] implemented to resolve all possible tactical elements of the position before the evaluation function is applied to the resulting positions.

Since the amount of positions that have to be evaluated grows roughly by factor 40 for each new level of depth, many heuristics and algorithmic strategies are necessary to cut unnecessary searches and evaluations. The main tool used for this by Stockfish is the alpha-beta algorithm. It is a variant of the minimax algorithm that is common for the evaluation of all kinds of two player zero-sum games. Since it is assumed that both players will act optimally, one cannot simply use the highest value of all the evaluations as the result. Instead, one has to go through all the options to check whichof the possible moves of one player leaves the worst options for the other player, since the other player will always choose the best of these options. The alpha-beta search is an optimization of this algorithm that allows eliminating the need to search through all of the different paths. Assume that we already have found a move that guarantees player 1 an equal position, i.e. there is no move from player 2 in response to this move, which leads to any advantage for player 2. From now on, when we consider alternative moves for player 1 and any of the evaluations show an advantage for player 2, we can skip the rest of the search for this particular move, since it is inferior to the move that we already found before. Clearly, this approach works best if we consider the best moves as early as possible. If we would consider the moves in ascending order of strength, we would have to go through all of the possible positions. If instead we always manage to consider the strongest move for each player first, the alpha-beta algorithm would enable us to reduce the nodes that have to be evaluated from x to something close to the square root of x. Therefore, move-ordering is another very important part of the algorithm. Once the move list is generated, it gets ordered using a number of heuristics, most importantly trying the best move from the previous search first, if the search has been deepened (Silver et al. 2018, Samuel 1959).

Apart from these very sophisticated algorithms, Stockfish uses an opening book to choose moves in the first phase of the game as well an endgame tablebase that includes the best moves for all positions with six or less pieces left on the board.

The above-described techniques are used by most of today's strong chess engines as well as by earlier versions such as Deep Blue.

3.2 AlphaZero

AlphaZero uses none of the techniques described above.

At the core of AlphaZero is a deep neural network that has been trained via reinforcement learning[3]. There is no domain-specific knowledge or data used as input; the

[2] Quiescence searches are part of traditional chess engines for a long time already. The challenge for the programmer is to extend the search until a position is reached that is suitable to be evaluated by the evaluation function but use minimal resources to do that. For different techniques, see for example (Beal 1990).

training phase has been done exclusively via self-play. The same approach has been used for the games of Shogi and Go successfully[4].

Chess moves are represented in a two-step process in AlphaZero. The first step is to pick up a piece, i.e., identifying the square from which a piece is moved. The second step is deciding which move this piece should make. The input into the neural network therefore contains the 64 squares and all the move possibilities each piece would have on each of these squares. Each square is treated in the same way; this leads to the coding of many illegal moves in specific situations, which are then masked out by setting their probability to zero, reducing the move possibilities to those that are available according to the rules in the actual position. For each square, there are 56 possible queen moves (these moves also code rook, bishop, king and most pawn moves), one to seven squares in all eight directions in which the queen can move. Additionally there are 8 knight moves and 9 possibilities for pawns to promote in a piece different from a queen, either by moving to the last rank or by capturing a piece on the last rank in one of the two diagonal directions. This adds up to an input of 8x8x73 equalling 4672 move possibilities in all positions. Additional inputs are needed to represent information about special rules: the castling rights, the number of repetitions of the present position and the move count without progress with respect to the 50 moves rule (Silver et al. 2018).

The output of the neural network is a vector with two numbers for each move. The first one signifies the percentage of search time, which was attributed to this move and its consequences. The second one represents the estimated win probability that the neural network assigns to this move. This means that the neural network has the function of the evaluation function in a classical chess engine. It is trained to assign a probability to each of the possible moves, which represents the win probability when making this specific move. The metric of evaluation is one of the most obvious differences to classical chess engines. This evaluation via probabilities can be interpreted as integrating two aspects, which are not well-represented in Stockfish's evaluation function: The general complexity of the position and the potential risks that result from choosing a specific move. Essentially, this way of representing pays tribute to the fact that the computer cannot calculate all possibly relevant lines and the resulting uncertainty. This will be larger in complex positions with more moves, which cannot be ruled out quickly[5].

[3] I cannot provide a description of reinforcement learning or machine learning in general in this paper. Goodfellow et al. (2016) provide a good introduction.

[4] In 2018 Leela Chess Zero, an open-source program with the same approach has been released. By now it has reached a level comparable to Stockfish.

[5]Two behaviours that can be observed in AlphaZero's play can be directly related to this modelling choice. 1) In positions with a large advantage, AlphaZero might simplify the position even on the cost of some of the advantage. While Stockfish would always search for the option that promises the largest advantage, there is no incentive for AlphaZero to win quickly or to achieve unnecessarily large material advantages. This might sometimes lead to moves that seem counter-intuitive for humans such as unnecessary underpromotions. 2) In positions, which AlphaZero evaluates as negative, it might try to complicate the position, making moves that still have a higher degree of uncertainty instead of a well-analysed move, which is evaluated with a very low success probability (Sadler & Regan 2019).

The training of the neural network has been done entirely by self-play over a timespan of nine hours. At the beginning of the training phase, all the parameters in the neural network were initialized with random values, leading to seemingly random moves by AlphaZero. Most of these games ended in draws because of the 50-move rule[6], but some of them were decisive and based on these games, the parameters in the neural network were adjusted. Repeating this process for nine hours and 44 million games, the millions of parameters in the neural network were adjusted repeatedly. Via the tuning of these parameters, AlphaZero can represent patterns and strategies. After each decisive game, the parameters in the network are updated to evaluate each of the positions of the game as better or worse depending on the outcome of the game (Sadler & Regan 2019).

We can conclude that there are two major differences in the architectures of AlphaZero and Stockfish. The evaluation of positions is approached with different metrics. Stockfish evaluates positions in pawn equivalents; AlphaZero assigns win probabilities to positions. The most important difference lies in how the evaluations are reached. Stockfish uses handcrafted domain-specific knowledge in its evaluation function; AlphaZero reaches the evaluation of the position through its neural network, which has been trained by self-play and without the input of any domain-specific knowledge except for the rules of the game.

4 Epistemic opacity

> *"Whilst we cannot understand exactly how AlphaZero is thinking, we can explore the ways in which AlphaZero generates its innovative and active plans, and how it conducts its ferocious attacks through analysing its games."* (Sadler & Regan 2019, 74)

It is certainly not appropriate to associate the calculations that lead AlphaZero to make a decision with thinking. If interpreted a bit more metaphorical, this quote gives a first answer to one of the main questions of this paper: What can we learn about how Alpha-Zero works and how is it different to what we can learn about Stockfish or similar other chess engines?

Since AlphaZero managed to beat Stockfish, it would be very interesting and potentially beneficial for our understanding of the game of chess, to learn on what AlphaZero bases its decisions. It seems to be a reasonable assumption that it must have something encoded about the game of chess that goes beyond the handcrafted domain-specific features, which have been encoded in Stockfish.

The thesis that I want to defend here is that the introduction of a deep neural network and machine learning led to a different kind of epistemic opacity in AlphaZero than the one in Stockfish. This new kind of epistemic opacity prevents us from confidently identifying on what AlphaZero bases its decisions.

[6] A game of chess ends in a draw, if for 50 moves in a row, no pawn is moved and no piece is captured.

4.1 Defining epistemic opacity

At first, it seems counter-intuitive that either Stockfish or AlphaZero can be epistemically opaque in any way. Both are based on algorithms, which are determined processes. For each of the calculations that are done in any of the chess engines, it is always clearly defined which rule or calculation has to be applied next. In principle, all of these calculations could be done by pen and paper or printed out.

Nevertheless, in the philosophical debate epistemic opacity is discussed very prominently. Especially in the context of software that is based on machine learning techniques as in the case of AlphaZero, their potential black box-nature is a widely discussed topic. The introduction of such methods into the scientific practice raises questions about the aims of science and necessity of explanations or understanding.

Let us take the most commonly used definition of epistemic opacity by Paul Humphrey as a starting point to understand what is typically meant by epistemic opacity and why it is useful discussing it in the context of computer models and machine learning.

> *"[A] process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process."* (Humphreys 2011, 139)

One of the potential sources of opacity is the sheer amount of calculations that are done by Stockfish or AlphaZero during a game of chess, which lead to the evaluations of positions and the decision for a specific move. Stockfish calculates up to 60 million positions per second, so it is quite clear that no human cognitive agent will ever go through all of the calculations that are made during one entire game for example. Even though AlphaZero calculates less positions per second (Silver et al. 2018), the amount of calculations still exceeds everything a human agent would be able to go through in a reasonable time. Regarding AlphaZero, one can additionally argue that the steps during the training phase are epistemically relevant elements of the process, since they determine the final values of the parameters in the neural network and thereby have a significant impact on the process. This adds a large number of calculations to the epistemically relevant elements[7]. It is therefore clear that the basic kind of epistemic opacity that Humphreys defines in the first sentence of his definition is present in both Stockfish and AlphaZero.

Are these opacities essential in the sense of Humphreys' definition? To argue for this, one has to show that it would be impossible for the cognitive agent to go through all of these calculations because of their nature. Let us take a human as the cognitive agent

[7] There is a debate about what should count as epistemically relevant. Durán (2018) for example points to the fact that a limited amount of information about an algorithm might be enough for the justification of results. In this paper, I follow the more widespread approach to count all parts of the process as epistemically relevant to it. The reduction of epistemically relevant elements requires background information. There is not enough of it available yet in this case, therefore using the wide approach seems to be more appropriate.

X. He is limited through his life span and if we assume a discrete amount of time that is needed for each of the calculations, we can determine a maximum number of calculations that one human could possibly do and thereby determine a threshold for this opacity to become essential. This approach is however not very informative about the nature of the source of opacity. It seems somewhat arbitrary that there should be some number of calculations x, which are necessary for a process, for which the opacity is not essential; if at the same time, adding just one more calculation would make the opacity essential.

As has been argued in Boge & Grünke (2020), there might be another useful differentiation between types of opacities, namely "contingent epistemic opacity" and "fundamental epistemic opacity". Both epistemic opacity and essential epistemic opacity, as defined by Humphreys, are contingent on the nature of the cognitive agent. Fundamental epistemic opacity on the other hand refers to opacities which are grounded in the nature of the process rather than the nature of the agent:

> *"[A process is fundamentally epistemically opaque, if] given any agent with any nature, at no time will the agent be able to obtain all relevant pieces of information about the process." (*Boge & Grünke 2020, 9)

A thought experiment shows that the opacity, which is caused by the amount of calculations, is contingent. Consider the Laplacean demon, understood as an entity with unlimited cognitive resources, which would be able to perform as many calculations as necessary without using up any time (de Laplace 1814). For it, the algorithmic transparency of the processes would be enough to render them epistemically transparent (Boge & Grünke 2020).

With respect to this source of opacity, AlphaZero has introduced no different kind of opacity.

4.2 Model-opacity as fundamental opacity

The definition by Humphreys shapes the concept of epistemic opacity as something that is connected to processes. This makes it natural to look for sources of opacities, which may arise in the process of coming to the results via a string of calculations or the training phase of the neural network.

Sadler & Regan (2019) focus on the question "how AlphaZero is thinking" (74), but in order to learn the new things, which AlphaZero might have discovered about chess that make it more successful than traditional chess engines, we need to understand *what* AlphaZero is thinking (thinking must be understood metaphorically again, of course). All of AlphaZero's decisions are based on the neural network that has been trained. By tuning the parameters of the neural network, AlphaZero represents features of the positions, similarly to the way features of positions are represented in the handcrafted evaluation function of Stockfish. In the case of AlphaZero however, it is not clear which features are actually represented in the neural network, since they are developed during the training phase and not preselected through the programming. It is possible that AlphaZero learns features, which have been chosen for Stockfish, such as safety of the king or pawn structure. It could also be the case that AlphaZero learns features, which are a lot more complex and far re-moved from the human way of describing concepts in chess. This uncertainty about what features are modelled in the neural network may constitute a different kind of opacity. It is not process-based but rather concerns the connection between the neural network and the real-world phenomenon, so it seems useful to call it *model-opacity*[8].

One natural approach of trying to overcome part of this opacity is a classical experimental one[9]. You come up with a hypothesis about the behaviour of the system and collect empirical evidence for or against it. The main source for empirical evidence in the case of AlphaZero are the games that have been published. Sadler & Regan (2019) analyse the games and try to identify features in AlphaZero's play. They show examples of well-known strategies, which have been adopted by AlphaZero, as well as positions in which AlphaZero does not follow any of the classical plans developed by humans and programmed into Stockfish[10]. However, the complexity of chess makes it impossible to isolate any specific feature in almost all cases. Each position can be described by many different features and their interactions. At the same time, these features cannot be deduced directly from the rules of chess but are human descriptions of patterns that have been recognized. The number of ways to describe features of a chess position is infinite and many might be functionally nearly equivalent but differ in some very specific ways. Therefore, in addition to the problem that we cannot isolate specific features, it is impossible to be sure that we have conceived all of the potential hypothesises. Even unlimited resources would not solve this problem. There is no algorithm that could solve this problem and therefore there is no algorithmic transparency, which would render this opacity contingent. Model-opacity therefore seems to be fundamental (cf. Boge & Grünke 2020 for a similar case).

4.3 Discussion of the opacities

The contingent opacity caused by the amount of calculation steps is similar for both Stockfish and AlphaZero. Even though there are differences in the concrete amount of calculations and AlphaZero has the training phase of the neural network, which influences the results and does not have a complementary part in the Stockfish architecture, the nature of these opacities is the same. All of these calculations are clearly defined and determined processes that can be solved if enough resources are available.

As described above, this is not the case for model-opacity. However, model-opacity does not exist for Stockfish. It does not exist because the modelling for Stockfish was intentional and goal-orientated. The modellers can be asked about the reasons and intentions for the implementation of specific features. This does not mean that Stockfish never makes moves which come unexpected to the modellers. No one who modelled parts of Stock-

[8] Model-opacity is discussed in Boge & Grünke (2020) for an example from high-energy physics. Sullivan (2019) uses the same term for the "complexity and black box nature of a model" (1). She introduces the term "link uncertainty" to describe the situation, that there is not enough empirical or scientific evidence to prove the connection between the model and a real-world phenomenon.

[9] Another approach is to find positions in which engines obviously misevaluate and try to derive information about the system from that (similar to adversarial networks in image classification). There are a number of constructions of positions with enormous material advantage for one player in which a human can quickly see that there is no way to achieve any progress but the computer nevertheless evaluates the position as very advantageous.

[10] The minority attack in the Carlsbad structure is used as a prominent example

fish has a complete comprehension of the interaction of all parts of the program and can calculate all the consequences of these interactions and therefore no one can make a confident prediction. Stockfish outperforms all human players in the game of chess after all. This is parallel to AlphaZero but can be explained with contingent opacity. What is possible when Stockfish makes an unexpected move is to analyse the reasons how the decision to make this specific move came about. By reviewing which evaluations led to the move and which features were evaluated in what way, a reconstruction of the decision process is possible. This is exactly what seems to have happened in the match between Deep Blue and Garry Kasparov. During the first game of the match, a bug in the code led to an unexpected move by Deep Blue. After the game, the programmers of Deep Blue could track down the reason why this move was chosen and fixed the bug (Anderson 2017). If this would happen with AlphaZero, the programmers would not know how to change the code of AlphaZero. There is no way to predict what the consequence of changing a single weight in the neural network would be for example. There is no way of confidently knowing about the way specific features are represented in AlphaZero's neural network and consequently they cannot be manually altered with a specific goal in mind.

This opacity of AlphaZero is of a different kind than the contingent ones and it does not only prevent humans from intentionally altering the program in an advantageous way but also prevents them from isolating and understanding the way in which AlphaZero represents chess, and which features of the game it chose in its neural network.

5 Conclusion

In this paper, I described the development of artificial intelligence in chess with its present-day culmination: the introduction of the neural network-based AlphaZero in 2017. I highlighted its differences from Stockfish, one of the highest rated traditional chess engines. The two most significant differences are the modelling process and the metric of evaluation, most notably the creation of the neural network of AlphaZero without any domain-specific knowledge about chess except for its rules.

The different modelling processes lead to different epistemic opacities. Both Stockfish and AlphaZero are epistemically opaque in a contingent way due to the very large number of calculations necessary during the training phase of AlphaZero as well as during the play phase for both Stockfish and AlphaZero. AlphaZero additionally also has model-opacity, which seems to be a fundamental kind of opacity. This opacity originates from the method that is used for creating the neural network: machine learning or, more specifically, reinforcement learning. Contrary to the modelling process of Stockfish, the modelling is not done intentionally by a human programmer, but through reinforcement learning – a statistical approach. This seems to make it impossible to connect the "reasoning" of AlphaZero to human reasoning in chess.

This fundamental opacity is not problematic in chess. Human chess players can still follow AlphaZero's games, use it as an inspiration and develop ideas and concepts from it, which they can integrate in their own games. For applications in domains other than games however, it is potentially ethically very problematic if human agents base their decisions on the output of a neural network with fundamental opacity.

Acknowledgements

The research for this paper was partially conducted as part of the research unit The Epistemology of the Large Hadron Collider, funded by the German Research Foundation (DFG; grant FOR 2063). I want to thank the participants of the Budapest Workshop on Philosophy of Technology 2019 for a very helpful discussion on the topics of this paper. I would also like to thank an anonymous reviewer for helpful suggestions on improvements of the original manuscript.

References

Anderson, Mark Robert. "Twenty years on from Deep Blue vs Kasparov: how a chess match started the big data revolution." Accessed May 12, 2020. http://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882

Beal, D Beal, Don F. "A generalised quiescence search algorithm." Artificial Intelligence 43, no. 1 (1990): 85-98.

Boge, Florian, and Paul Grünke. "Computer Simulations, Machine Learning and the Lapla cean Demon: Opacity in the Case of High Energy Physics", forthcoming in Resch, Kaminski, and Gehring (Eds.), *The Science and Art of Simulation II*, Springer (expected 2020).

Chabris, Christopher. "The Surprising Return of Odds Chess." Accessed May 12, 2020. https://www.wsj.com/articles/the-surprising-return-of-odds-chess-1461339115

Chessentials. "History Of Chess Computer Engines." Accessed May 12, 2020. https://chessentials. com/history-of-chess-computer-engines/

de Laplace, P. S. *A Philosophical Essay on Probabilities*. London: Chapman & Hall, Ltd. Translated from the 6th french edition by Frederick Wilson Truscott and Frederick Lincoln Emory, 1902 [1814].

Durán, J. M. 2018. *Computer Simulations in Science and Engineering: Concepts—Practices— Perspectives*. Cham: Springer Nature.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Humphreys, P. "Computational science and its effects". In Carrier, M. and Nordmann, A., editors, *Science in the Context of Application*, 131–142. Dordrecht, Heidelberg: Springer, 2011.

Morton, Ella. "The Mechanical Chess Player That Unsettled the World." Accessed May 12, 2020. https://slate.com/human-interest/2015/08/the-turk-a-chess-playing-robot-was-a-hoax-that-started-an-early-conversation-about-ai.html

Sadler, Matthew, and Natasha Regan. 2019. *Game Changer*. New in Chess.

Samuel, Arthur L. "Some studies in machine learning using the game of checkers." *IBM Journal of research and development* 3, no. 3 (1959): 210-229.

Schwartz, Oskar. "Untold History of AI: When Charles Babbage Played Chess With the Original Mechanical Turk." Accessed May 12, 2020. https://spectrum.ieee.org/tech-talk/tech-history/dawn-of-electronics/untold-history-of-ai-charles-babbage-and-the-turk

Shannon, Claude E. "XXII. Programming a computer for playing chess." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41, no. 314 (1950): 256-275.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science* 362, no. 6419 (2018): 1140-1144.

Standage, Tom. 2003. *The Turk: The Life and Times of the Famous Eighteenth-Century Chess-Playing Machine*. New York: Berkley Trade.

Sullivan, Emily. "Machine Learning and Understanding". forthcoming in: *British Journal for the Philosophy of Science* (2019). Available at: http://philsci-archive.pitt.edu/16276/

Thicknesse, Philip. 1784. *The Speaking Figure, and the Automaton Chess-player, exposed and detected*. London: Stockdale. 8vo. pp. 20.

Author information
Paul Grünke, Karlsruhe Institute of Technology, https://orcid.org/0000-0002-3576-1921

A black market for upvotes and likes

Mihály Héder

Abstract

This article investigates controversial online marketing techniques that involve buying hundreds or even thousands of fake social media items, such as likes on Facebook, Twitter and Instagram followers, Reddit upvotes, mailing list subscriptions, and YouTube subscribers and likes. The findings presented here are based on an analysis of 7,426 "campaigns" posted on the crowdsourcing platform microworkers.com over a 365 day (i.e., year-long) period. These campaigns contained a combined 1,856,316 microtasks with a net budget of USD 208,466.

Keywords: Crowdsourcing, facebook, twitter, youtube, reddit

Introduction

This article analyses marketing campaigns that have been executed through the hiring of *freelancers* or "*microworkers*" to complete short, menial tasks called *microtasks* that usually pay less than one US dollar each, and most often only around ten cents. These tasks include watching, liking, upvoting, and "+1"-ing items on web platforms featuring social media functions, like Facebook, Twitter, Reddit, and Instagram. A job description showing what such a campaign looks is given below (an actual, observed example):

Title: Facebook Like: <REDACTED String>
Payment: USD 0.15
Number of workers accepted: 100

Job description:

WARNING: we manually review almost all of the submitted tasks. Thus, if we find that you have ignored the instructions (i.e., posting on a wrong site, using non-unique or nonsensical content), you will be permanently banned from our system. You must have 50 Facebook friends.

1. **Go to <REDACTED URL1>**
2. **Visit the URL shown at <REDACTED URL1>**
3. **On that page, you will see a Facebook Like button. Click on it**
4. **Submit your Facebook profile URL on <REDACTED URL1> to verify that you have completed the task (make sure you have set your Facebook profile to Public**
View in order for your task to be verified) to get a 7 character confirmation code (...)

Campaign example 1: Buying facebook likes

Completing this task pays USD 0.15 for the microworkers. In this case, one hundred freelancers were sought to do the task.

There are specialized web platforms for the brokering of small tasks. One of them is microworkers.com, which was investigated for this article. This platform lets employers run campaigns for a fee, usually 10% of the payment made to the freelancing internet users who do the job, called microworkers. Therefore, the 100 Likes above would cost the client posting the job only USD 16.5. The platform has hundreds of thousands of microworkers (Nguyen 2014).

The platform was not specifically created for promotional or marketing purposes, and indeed data processing, survey-based research, and software testing jobs are also posted on the platform. However, as shown below, the majority of tasks can be categorized as being of a promotional nature (See Table 3).

1.1. Article Structure

The structure of the present article is as follows. After this introduction, the terminology is presented and a number of ethical questions are posed. This is followed by an outline of the related work in this field. The next chapter describes the research aims and details of the observation, as well as the limitations of the research. This is followed by the Results section, which also contains a number of separate subchapters covering various aspects of the investigation into the most important gray promotional activities, including estimates about their efficiency, security concerns, and possible counter-measures, where applicable. The Results section also describes some observed campaign-supplementing techniques. Finally, the Conclusions section offers a summary and outlines some possibilities for future work.

2. Why call it a black market?

In the opinion of the author of this paper buying likes, followers, votes, upvotes, retweets, etc. (the generic term for these used in this paper is social media activity) for promotional purposes is an unethical practice, therefore we use the term “black market” to describe this part of the industry. This is not to say they are necessarily breaking any rules or laws—that question is outside the scope of this paper.

This kind of activity is considered unethical for the following reasons. a) The users of a social platform are misled by a page or post having an artificially inflated number of likes, followers, etc. Normally, it is impossible for users to differentiate between paid activity and genuine activity. The conventional semantic behind a like, upvote, or follow is a statement of approval of the content in question. In other words, the microworkers are paid to lie in the sense that they are paid to pretend to like/endorse/approve content that they most likely have not truly read or watched. b) The microworker’s “employment” is arguably rather exploitative, because of the low payment and the apparent lack of any powers against the clients (see table 5). c) Content creators and social media users who do not employ such practices are clearly disadvantaged. Finally, d) it is easy to see that if paid social media activity were more universal, it would create an unsustainable social media environment.

Besides paid social activity, there are other highly controversial campaigns based around social media. One such typical activity involves creating accounts on behalf of a client and then the microworker handing over the user name and password to the client. An example of this kind of activity is given in the following (actual, observed example):

Title: Gmail: Create an Account
Payment: 0.16

Number of workers accepted: 230

Job description:
Note: I will Check American Name and Profile Picture otherwise I have to decline you.

1. **Go to www.fakenamegenerator.com**
2. **Choose proper American name**
3. **Go to Gmail.com**
4. **Create a new Gmail account using details from www.fakenamegenerator.com**
5. **Upload a good profile picture in Gmail**

For proof:

- **Give Password**
- **Give a Recovery Mail so that I can change it later**

Note: Make sure all are perfect otherwise, I will decline your payment.
Required proof that task was finished?

1. **Gmail and Username**
2. **Password**
3. **Recovery email**

Campaign example 2: Acquiring gmail accounts

In this case, the client was able to acquire 230 Gmail accounts for a mere $ 40.48. We might speculate that these accounts (and similarly those created for YouTube, Twitter etc.) will be used for promotional purposes, controlled by the client, while also raising all the ethical concerns highlighted in the previous example. However, this speculation might actually be optimistic. Fake accounts like these could also be used for more sinister purposes, such as as tools for spreading fake news (Allcott und Gentzkow 2017) for political purposes or for committing fraud.

To sum up, it seems justified to state that there is a black market for "fake" accounts and social media activity and that promotion done using this market represents an ethically gray area.

The author wants to point out that this is not to say that microworkers.com or any other platform is in itself immoral or designed to be a black market—such platforms also offer a useful venue for valid projects, like data processing at scale, acquiring subjects for survey-based or interactive online scientific research, monitoring a competitor's public online activity, or counting objects on images - several such campaigns were observed.

3. Related Work

The marketing techniques discussed in this article are related to techniques called sock puppetry and click farming. Sock puppetry means the control of many social media accounts by one person or a small group of individuals. A report on such activity was published in The

New York Times (Caldwell 2007). The method of control is a crucial difference in these efforts. As we will see some in Section 6.5 clients buy hundreds or thousands of accounts that they can use themselves. In this case there are technical possibilities to detect the puppetry, by noticing when a very high number of users are logging in from the same network location or use the browser client fingerprint (Laperdrix u. a. 2016).

But the method of control can also be an order from a client to a cohort of users to perform some activity (but the client itself never logs in to any of their accounts). In this case detection of the activity is much harder if the client takes some precautionary steps (see Section 6.5). Sometimes this is called meet puppetry (Cook u. a. 2014) referring to the fact that it involves real freelancers.

Click farming is another related term. Click farms are actual workplaces in developing countries where a large number of employees are performing short task sometimes for as low as 1000 likes for 15 USD(Arthur 2013). In current reporting these are often called troll farms (Smith 2018).

The platform investigated by this article, microworkers.com differs greatly from a click farm as it is a completely distributed crowdsourcing tool, but some of the campaigns done here might be similar to those done by a click farm.

On the deceptiveness of such campaigns in comparison with traditional advertising (where the prospective customer is aware that it is being presented with an ad) is well described by Del Riego (2009) and Forrest und Cao (2010) in connection with then-new US Federal Trade Commission guidelines on endorsements and reviews.

This article focuses on probably way smaller market available on microworkers.com, which is, however, easily accessible for freelancers and is not specialized to gray marketing in particular. In contrast with the services that directly offer followers and social media activity (Fiverr, SeoClerks, InterTwitter, FanMeNow, LikedSocial, SocialPresence, Social- izeUk, ViralMediaBoost (De Micheli und Stroppa 2013)), here the client has to organize its own campaign and orchestrate the freelancers on the crowdsourcing platform. This allows for creativity and innovations in the campaign methods.

Nguyen (2014) explained the idea behind microworkers.com, founded in 2009, as well as reported its user count at the time of writing (presumably 2014). Howe (2006) also reported on microworkers.com as a crowdsourcing platform.

According to Nguyen, the platform had over 600,000 users from 190 different countries. The aim of the microworkers.com as a project was to aid brokering crowdsourcing campaigns. As the article explained,

> *In crowdsourcing platforms, there is perfect meritocracy. Especially in systems like Microworkers; age, gender, race, education, and job history do not matter, as the quality of work is all that counts; and every task is available to Users of every imaginable background. If you are capable of completing the required Microtask, you've got the job.*

The campaign templates on the landing page of the platform are great sources of inspiration for what could possibly be achieved through crowdsourcing: participating in market research, captioning documents and video, categorizing images, testing websites and applications, and so on. The fact that the majority of public campaigns visible on the platform are mostly employing controversial promotion techniques does not seem to be the result of the platform design or intentions.

Hirth u. a. (2011) investigated microworkers.com in order to compare it to the much better understood Amazon Mechanical Turk (Paolacci u. a. 2010, Buhrmester u. a. 2011). They correctly identified a main difference between the portals: the payment mechanism. At the time, it was basically impossible to use MTurk without a US-based credit card, while the microworkers website allowed payments to be made with Moneybookers (called Skrill today). This helps explains why the author of this paper and possibly other non-US- based researchers first discovered microworkers. Works by Gardlo u. a. (2012) and Crone und Williams (2017) aimed to assess the usefulness of the platform for scientific purposes; and indeed, scientific projects regularly, though relatively infrequently, appeared on the platform. However, it is possible that this difference in payment methods is only one of the reasons behind the nature of the campaigns conducted on each.

Hirth et al.'s work (Hirth u. a. 2011) indicated that gray promotional campaigns were already existed as long ago as 2011: "Signup", "Click or Search", "Voting and Rating" were already featured as campaign categories; however, the payments offered were slightly higher than today.

The connection between social media and marketing was analyzed by Thackeray u. a. (2008) as early as 2008. Of course this work concentrated on the legitimate social media strategies firms might embrace, such as paid search results, where the brand buys a presence in the search results. As Yang und Ghose (2010) explained, these are usually placed in a separate area on the results page, together with being clearly to indicate that they are paid for or they may be labeled as an ad, e.g., on Facebook (the difference between "paid" and "organic" (showing up in non-paid results) links is less emphasized in today's search engines but remains clear). Rutz und Bucklin (2011) demonstrated how tying paid search results to generic search terms might increase the success of a branded paid search.

The search and engage campaigns (see section 6.2) discussed in the present paper are different. They don't try to increase visibility by buying paid results. These represent the dark side of search engine-based advertising, whereby they try to directly manipulate the organic links.

It was envisaged (Zhang u. a. 2013) that the identification of the key influencers on social media could be crucial for effective viral marketing—but with the gray marketing techniques presented here, influence is attempted to be created directly, albeit artificially.

It was also envisioned that customer-generated content on blogs, etc. would be crucial for promotional activities—in the present paper, campaigns seeking to manufacture legitimate-looking customer content are analyzed. In other words, these campaigns, albeit unethical, are sometimes the effective counterparts of hard-to-operate social media marketing tactics, or in other words, they are controversial shortcuts to followers and likes and brand-friendly social content.

Confessore et al.'s recent work reported on in the New York Times (Nicholas Confessore und Hansen 2018) covered a very similar theme to this article, but was focused mostly on Twitter and on the activities of a company called Devumi.

4 Aims and Methods

4.2 Research aims

The primary aims of the present research were to identify the different schemes employed in microworker-based gray marketing on microworkers.com and to then categorize them, attempt to uncover how they fit in a wider strategy, estimate their limits and effectiveness, and to utilize this knowledge to provide general insights into these kinds of campaigns.

Second, the scale and typical budget of these campaigns, and their relative share of the overall activity on the microworkers.com platform were also measured and reported herein.

4.2 Observation of campaigns

This research project attempted to observe all campaigns posted on microworkers.com from 22 February 2016 to 21 February 2017. The site was checked several times a day during this period. In total, 7,426 campaigns were observed during the period. Each campaign was manually categorized and the aggregate numbers of categories (payments, number of tasks) were updated.

4.3 Categorization

The campaigns were categorized in terms of two dimensions: the related target platform (Facebook, Google Plus, SoundCloud, Twitter, etc.), and the specific activity (search and engage, like, comment, sign up, etc.). The categorization was based on the campaign title and text and proved to be quite straightforward as the names of the platforms were clearly stated in the title and represented unambiguous brand names, and as the activity to be performed was almost always explained in an itemized list in the posting. Obviously certain activities were further linked to certain platforms (like retweets can only be done through Twitter), but others, like search and engage can be done on multiple platforms.

The platform labels utilized are summarized in Table 1. while the activity labels are summarized in Table 2.

Also, each campaign could belong to multiple activity categories, by using multiple tags.

Platform categories

Ali (Alibaba, AliExpresS)	**Instagram**
Amazon	**Bing**
MBS: (microblog instant share): blog- ger.com, Pinterest, Digg, Tumblr, 9gag or other blogs or quick sharing platforms	**Mix**: (SoundCloud, Mixtape, datpiff)
Browser add-on (e.g., Chrome extensions)	**RDT** (a traffic generator site, the name of which is redacted from this article)
Other: (500px, Wordpress.com, Snapchat, Skillshare, Hotmail, LinkedIn, Coursera, Bitbucket, Snapchat, Steam, Yandex, other uncategorized)	**Question** sites: (Yahoo Answers,Quora)
eBay	**Reddit**
Forum (Disqus, Warrior Forum, other forums)	**Redacted**: (not visible in description because of the rotator technique, explained below)
Facebook	**Smartphone** (iOS and Android)
Gmail	**Twitter**
Google (search)	**Yahoo** (search)
Google+	

4.4 Limitations

On microworkers.com, there are invite-only campaigns as well. Unfortunately, there is no information freely available on these campaigns or their share of the total number of campaigns. The nature of invite-only campaigns, for which clients apparently hire tested and trusted microworkers, could be a subject for further research.

Some campaigns might have slipped trough between two observations, meaning that their full life cycle very short (only a couple of hours). This does not appear to be typical but cannot be ruled out. Therefore, it can be said that in reality there were possibly more than 7,426 public campaigns and an additional, unknown number of invity-only ones.

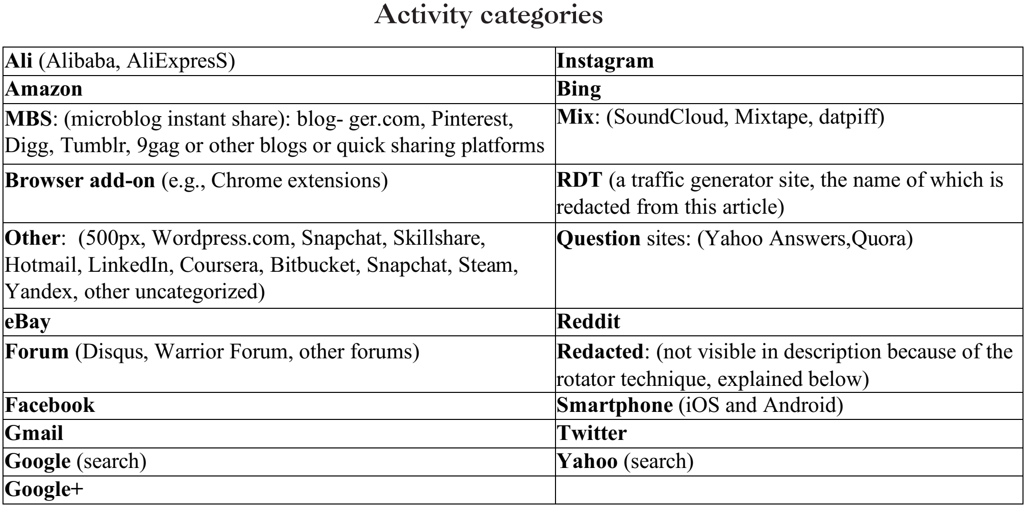
As explained above, the campaigns were manually categorized by the author. This categorization, because it relied on objectively observable features of the campaigns, did not require significant subjective judgment. Therefore, in the context of intersubjectivity, it should not be a serious limitation that there was no multiple-person cross-checking performed for interpretation of the category labels.

Finally, there are obviously other brokering platforms for such microtasks, but these are outside the scope of this investigation (in fact, some of those platforms seem to be using microworkers for recruitment purposes).

The author is confident that these limitations do not prevent the work from meeting its stated aims, as it is an explorative rather than exhaustive description of techniques and strategies.

4.5 Anonymization

Generally, all data presented in this paper (mostly microtask descriptions) is anonymized. Most of the actual URLs, person and company names, and other named entities are replaced with the string < *REDACTED(...)* >. When there are several URLs or names within



Activity categories

Ali (Alibaba, AliExpresS)	**Instagram**
Amazon	**Bing**
MBS: (microblog instant share): blog- ger.com, Pinterest, Digg, Tumblr, 9gag or other blogs or quick sharing platforms	**Mix**: (SoundCloud, Mixtape, datpiff)
Browser add-on (e.g., Chrome extensions)	**RDT** (a traffic generator site, the name of which is redacted from this article)
Other: (500px, Wordpress.com, Snapchat, Skillshare, Hotmail, LinkedIn, Coursera, Bitbucket, Snapchat, Steam, Yandex, other uncategorized)	**Question** sites: (Yahoo Answers,Quora)
eBay	**Reddit**
Forum (Disqus, Warrior Forum, other forums)	**Redacted**: (not visible in description because of the rotator technique, explained below)
Facebook	**Smartphone** (iOS and Android)
Gmail	**Twitter**
Google (search)	**Yahoo** (search)
Google+	

Table 2: Categories of activities

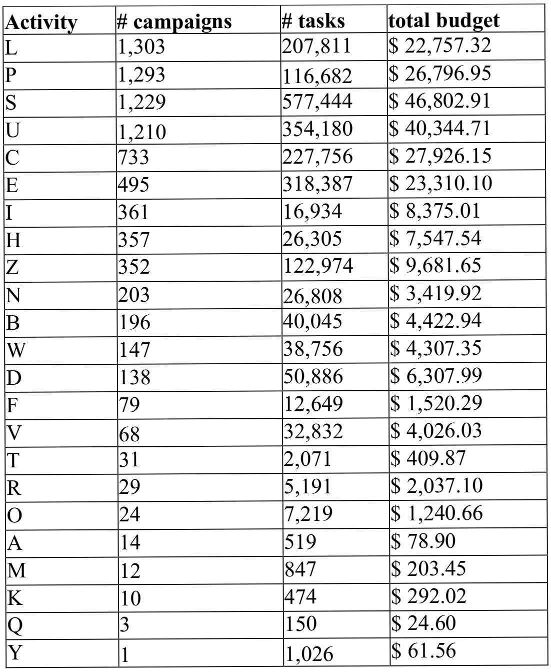
one example though, they are replaced with their own unique label so that they are not mixed up. Other than this modification, the job descriptions are copied herein verbatim.

Obviously, some basic URLs like Facebook.com or fakenamegenerator.com, for example, are kept because they are reported only for uncovering the campaign method but not its content, and also because the job descriptions would not be as understandable without them.

The reason for the redaction is that the persons, websites, and Facebook accounts mentioned in these task descriptions may be unwilling targets of a campaign. It is also probable that the customers of promotion campaigns are often not aware or may even have been misled about the methods employed on their behalf.

5 Results

Based on the observations, the following summaries were created.
Table 3 contains budget summaries by activity. The columns represent the activity code, number of campaigns, number of tasks, and net budget (without the 10% fee). The table is ordered by the descending number of campaigns.



Activity	# campaigns	# tasks	total budget
L	1,303	207,811	$ 22,757.32
P	1,293	116,682	$ 26,796.95
S	1,229	577,444	$ 46,802.91
U	1,210	354,180	$ 40,344.71
C	733	227,756	$ 27,926.15
E	495	318,387	$ 23,310.10
I	361	16,934	$ 8,375.01
H	357	26,305	$ 7,547.54
Z	352	122,974	$ 9,681.65
N	203	26,808	$ 3,419.92
B	196	40,045	$ 4,422.94
W	147	38,756	$ 4,307.35
D	138	50,886	$ 6,307.99
F	79	12,649	$ 1,520.29
V	68	32,832	$ 4,026.03
T	31	2,071	$ 409.87
R	29	5,191	$ 2,037.10
O	24	7,219	$ 1,240.66
A	14	519	$ 78.90
M	12	847	$ 203.45
K	10	474	$ 292.02
Q	3	150	$ 24.60
Y	1	1,026	$ 61.56

Table 3: Budget Summary by Activities

If we take out the category data processing, research participation, captcha solving, testing, installing and “other”, the remaining activities are purely for promotional purposes. This leaves 1,665,138 tasks, or 89.7% of the whole. It should be added that many of the install tasks appear to be promotional (see section 6.4). Counting these in would make the figure even higher. However, since for many such campaigns this aspect is impossible to tell, they are left out.

A similar summary for the platforms is given in table 4. The first column is the platform name, the rest is the same as before. The table is ordered by the descending number of campaigns.

From these values, the average payment for tasks related to certain activities and platforms could be calculated. Here is the distribution of the payments (not equal ranges):

The lowest paid wage was $0, and this was incidentally the biggest campaign with 99,999 tasks. The task was a simple visit to a link. The client promised future tasks for those who completed the task. More detail is provided on this task in section 6.5.

The highest paid wage was $3 for a task, but as can be seen from the above table, there were only 531 jobs offering between $1.1 and $3, while there were over a million jobs offering between 5 and 10 cents, making the higher paid tasks very rare indeed.

6 Campaign types and their analysis

6.1 Voting

Participation in voting (V) is a recurring activity on microworksers.com, with 68 campaigns posted featuring an aggregated 32,832 votes purchased. The top two voting campaigns seemed to be promoting a product and a sports team (2,130 and 2,000 tasks), the third was

Platform	**# campaigns**	**# tasks**	**# total budget**
Smartphone	1102	231,892	$ 27,702.07
Google+	823	57,203	$ 18,960.50
Other	797	349,937	$ 31,333.06
Twitter	747	83,731	$ 14,047.21
Facebook	666	115,967	$ 15,657.86
YouTube	616	248,914	$ 29,322.28
Reddit	539	77,104	$ 5,434.15
Redacted	501	132,401	$ 13,397.94
Google	330	261,923	$ 18,672.37
RDT	296	80,844	$ 8,034.94
MBS	218	38,859	$ 5,497.36
Instagram	172	19,421	$ 2,072.35
Amazon	156	55,760	$ 6,934.85
Mix	144	11,803	$ 1,458.55
Question site	126	16,190	$ 2,447.59
Forum	84	7,558	$ 964.15
Gmail	50	10,275	$ 2,908.35
eBay	22	6,014	$ 570.52
Yahoo	21	45,686	$ 2,355.13
Ali	9	3,945	$ 551.91
Bing	4	675	$ 69.00
Browser add-on	3	214	$ 73.96

Table 4: Budget Summary by Platforms

a giveaway voting for an expensive trip for couples, where the entrants were supposed to vote on the videos they uploaded about themselves. One entrant purchased 1,703 votes (the wording reveals that they bought the vote for themselves personally) for $0.12 each:

Title: Video: vote
Payment: 0.12

Number of workers accepted: 1703

1. Go to <REDACTED URL>

Give the video a VOTE by clicking on the heart below the video Required proof that task was finished?

1. Tell me how many votes I had after you've voted

Campaign example 3: Buying votes

Unfortunately while this can be seen as being clearly unethical, for a little more than $ 200 it could have been economically viable if the entrant won the vote. And it is even plausible that a local vote could have been won with just 1,703 extra votes bought.

Other votes were for titles like best auto repair shop, best bakery or "tradie of the year in Australia". There was also a census on how many New York City residents wanted to go on a date with a certain model. There were votes on temple photography, the best fintech firms, music mixes, the best female vocalists, several contests about the ranking of attractive persons, a vote on XXL Magazine, a vote on the best local charter flight provider, and so on. Some of these were clearly promoting a product or a performer or artist; others seemed to be clearly what we will call vanity-promotion.

The purchased votes ranged from dozens to about 1,000 at a time. It is hard to assess the overall efficiency of such campaigns, but it can certainly be said that for local contests,where the maximum number of people voting is expected to be measured in hundreds or thousands, it is very easy to rig contests this way, as it would cost only a few dollars.

task payment	total number of tasks
$0	99,999
$0.05-$0.1	1,059,172
$0.11-$0.2	581,146
$0.21-$0.3	63,904
$0.31-$0.5	35,893
$0.51-$1.0	15,671
$1.1-$3.0	531

Table 5: Number of tasks by payment range

6.2 Search and engage tasks

The common feature of these kinds of microtasks seems to be an attempt to manipulate the search results in search engines like Google, Yahoo, Bing or the search feature of Facebook. In most cases, a certain item is promoted, but in some rare cases, the goal actually seems to be to push unwanted result items back in the result list.

These campaigns appear to assume that search engines learn: if for given search term, a high number of users click on a particular result item, then that result item must be a good result for the search and therefore it will be listed early in the results listing. While the actual algorithms search engines use are proprietary and unpublished, it is known how they work in theory (Buttcher u. a. 2016). It is thus plausible that they can be tricked in this way to a certain extent. We know that user behavior is taken into account in Google, for instance, as (Clark 2015) reported that Google's novel AI solution, BrainRank, was the third most important factor (the technical term is "signal") when ranking pages. We also know that it learns from user behavior, hence it is plausible these systems can be tricked through paid user behavior simulating genuine interest.

A search and engage campaign therefore hires a large number of microworkers for searching the given terms and clicking on the promoted item in the search results.

An example of this type of campaign is given below:

Title: Google search
Payment: 0.08

Number of workers accepted: 1600 Job description:

1. **Open up Google.com (please use US version)**
2. **In Google, please search for this phrase: <REDACTED PERSON NAME>**
3. **Please click on any of the red boxed links you see in the attached file <the file is a Google result list screenshot, red rectangles designate what result items need to be promoted>**
4. **Stay on page for 1 minute (...)**

Campaign example 4: Google search

The top ten search and engage campaigns have the number of tasks offered as between 2,941 and 6,770. However, many of these campaigns seem to be part of the same project, making the biggest projects around 10,000-20,000 tasks.

An interesting tendency is that in many of these promotion campaigns, it seems that there is no marketed product involved, rather it is individuals concerned with their online persona who are the payers, in what is really another example of vanity-promotion.

The biggest search and engage project, with above 10,000 tasks, involved the promotion of a USA business executive's Wikipedia entry, whose name happens to be the same as a famous USA American football player and also a former USA congressman. The project must have been a success as currently the promoted page comes out top in a Google search when searching for that name. Naturally, it is impossible to establish the causal relation between the campaign and the current ranking with any certainty, especially this long after the campaign.

6.3 Social media activity

Paid social media activity involves task like creating Pinterest Pins, upvoting in Reddit, YouTube, or Google+, liking in Facebook, using Digg, Twitter, or Instagram, commenting on forums, upvoting on SoundCloud, Mixcloud or Datpiff, and so on. In terms of activity codes, this section covers B, C, F, H, L, N, T, W.

Campaign example 1 is in this category. The campaigns are usually straightforward and easy to do, therefore the payment is usually very low. Clicking on like, upvote, etc. are the lowest paid tasks. For instance, the 77,104 Reddit upvotes purchased during the 365 days study period cost less than $ 5,500 in total (see Table 4). The highest paying jobs in this category were those that require writing content that meets a set specification, e.g.:

Title: YouTube: Comment 3x (1-3)
Payment: 0.30

Number of workers accepted: 90 Job description:

1. **Go to www.youtube.com/channel/<REDACTED>.**
2. **Post a positive relevant comments on the videos found in first three links Important: Comment must be at least 10-15 words long and cannot be generic and must include the following words <REDACTED Person name> and the word "Florida".**
3. **Stay on each video page for 1 minute**

Required proof that task was finished?
1. **YouTube display name**
2. **Copy of the comments you've posted**
3. **URL to YouTube videos where comments were posted**

Campaign example 5: Youtube comments

However, in other cases the freelancer is asked to copy-paste the comment content, and the job then pays less:

Title: YouTube: Comment 3x (<REDACTED>)
Payment: 0.12

Number of workers accepted: 440 Job description:

1. **Go to the instruction page: <REDACTED URL>**
2. **Search Youtube.com for the key phrase**
3. **Copy-paste the supplied comments onto relevant video**
4. **Repeat steps 2 and 3 for 2 more videos (3 total)**

Required proof that task was finished?
1. **Your YouTube name**
2. **The search phrases you used**
3. **URLs of the 3 videos you commented on**

Campaign example 6: Youtube comments - the copy-paste method

For commenting, the most prominent platforms are YouTube (364 campaigns; 197,735 tasks, some paying for three comments), Instagram (101; 4,670), questions sites (81; 10,282), Facebook (30, 2,485), and a long tail of other forums (see some under Platforms; 118 campaigns).

Following a given account is done on Google+ (317 campaigns; 21,228 tasks) Instagram (54; 8,960), Twitter (20; 2325), Quora and Yahoo Answers (4; 466), and Google+ (1; 898). In must be noted that for Twitter, the clients often require the presence of some features from the freelancer, e.g.:

(...) To do this task, you need to have a Twitter account that meets the following requirements:- At least 100 followers - Your follower count needs to be at least double your following count (meaning - if you are following 100 users, you need to have at least 100*2 =200 followers) - The majority of the 20 most recent tweets are in English - At least 8 out of 20 most recent tweets have no links, are Not Retweets, and sound natural and interesting.(...)

Campaign example 7: Twitter account quality rules

This is obviously requested in an effort to imitate a real Twitter user and to not seem like a newly created one. These tasks pay bonuses too, meaning that the pay can reach as high as $ 0.25.

Posting (P) and Tweeting (T) were grouped together for being very similar. P and T is most prevalent on Twitter (716 campaigns; 80,211 tasks) and Google+ (485; 28,819). Connecting as a friend is mostly done on Facebook (13 campaigns; 9,190 tasks). Liking/Upvoting (L) is an activity performed on Reddit (539 campaigns; 77,104 tasks), Facebook (485; 71,200), Mixcloud and SoundCloud and Datpiff (143; 11,773), YouTube (63; 23,993), Instagram and Google+ (both 13 campaigns, 5,366 and 5,401 tasks, respectively), and on some other platforms (49).

The 10 biggest like/upvote campaigns ranged between 1,830 and 7,417 offered tasks (here, Facebook, Reddit, Instagram, Google+ campaigns were all in the top 10). The content of these top campaigns unfortunately were redacted using the rotator technique (see later in the article), but some of the remaining involved promoting persons not noted on Wikipedia (vanity-promoting); and some niche products. Size seems to be a limitation again, just like with search and engage campaigns: for celebrities with hundreds of thousands of followers, even as much as 7 thousand new likes hardly seem to matter, and the promotion technique does not seem to scale up to higher numbers.

This limitation might be not there in the case of comments (C) campaigns. The biggest comments campaign was conducted on YouTube, very similar to example 6, and it involved 21,140 tasks, three comments each, yielding more than 63,000 paid comments. The tenth-biggest involved 2,425 tasks, again three comments each. While a similar amount of likes would still represent just a fraction when it comes to comparing it to the likes received for the most popular YouTube videos, Facebook accounts, etc. For comments, the case is different, because only a small fraction of readers/visitors make comments. A cursory, non-representative investigation of the many YouTube videos reveals that it is very hard to find videos that have less than 20x viewers than comments. Thinking the other way around, 63,000 tendentious comments would suggest representing well over 1.2 million viewers, hence distorting the perception of actual viewers on what other's opinions are in relation to the topic.

However, in local communities with a smaller overall size, it seems that even small (L) campaigns can make a difference among the competition. Consider this example related to warriorforum.com (its self-description is: “The world’s largest Internet Marketing Community and Marketplace.”)

Title: Warriorforum Post: Comment Payment: 0.10
Number of workers accepted: 30
Job description:
Must have a Warriorforum account at least 2 months old, or a very active account if you joined recently.

1. **Go to: www.warriorforum.com/<REDACTED>**
2. **Post a comment/testimony relating to the post Must be original/no copying others post**

Tip on what you can post:
I learned a lot from the webinar,
I just join and I am excited, cannot wait to start working with <REDACTED>, this is an awesome program, etc.

Required proof that task was finished?

1. **Name used to leave comment**

Campaign example 8: Comments on a small forum

A cursory look at Warrior Forum reveals that the typical number of upvotes is around 10 and the number of comments (reply-s) is similar. The job offer above for 30 comments in example 8 would thus propel the entry among the top, while costing a mere $ 3.3 for the client.

6.4 Smartphone apps

A distinct area of paid activity is installing smartphone applications, using/testing and rating them. There were 361 such campaigns with 16,934 tasks worth $ 8,375.01. A typical campaign looks like this:

Title: Android App Testing (<REDACTED>): Download+Install+Honest Review
Payment: 0.50

Number of workers accepted: 370 Job description:

1. **Install the app below: play.google.com/store/apps/<REDACTED>**
2. **Download the app**
3. **Open the app for 30 seconds and test**
4. **Rate the app**

Optional: Leave 3, 4, or 5 stars

5. **Write an honest review in the Microworkers proof box only**

Required proof that task was finished?

1. **Your Google username**
2. **Paste your review**

Campaign example 9: Smartphone app ’’testing”

We can interpret this campaign in several ways. The charitable interpretation would be that this is an honest test for the software. Even though the app is only required to run for 30 seconds, the freelancers would open it on dozens of different Android devices, with different capabilities, screens, API levels, etc. Thirty seconds is enough to run some self-assessment and report back to a server. This test could thus have some value from a Software Engineering perspective and may be a valid assignment.

However, this kind of test is surely already long overdue in the case of a published application. Issues around crashing apps and unwanted startup behavior should have been resolved way before then. The author speculates that these kinds of campaigns are instead promotional. The clients are usually careful not to explicitly order five stars or positive reviews (albeit there are counter-examples of such). Yet, in such a campaign, an initial, visible user base is created for the app. Again, we can assume that this is more useful in niches than in competition with mainstream applications, as the leading apps have millions of users already.

It must be mentioned that this kind of microtask carries security risks for the freelancer and for the general public. The fact that the freelancer installs apps for a fraction of a dollar presents an obvious opportunity for breaching their smartphones. Although the current number of tasks in these campaigns does not seem to be high enough, in theory it would be possible to create a zombie network for DOS attacks or similar purposes.

6.5 Signup

A very common type of campaign is the signup (U) campaign. These campaigns involve creating an account meeting some client-specified requirements. In many cases, the microworkers will be required to hand over the account credentials. Example 10 below was one of the biggest signup and account handover campaigns observed.

Title: YouTube: Create an Account
Payment: 0.10

Number of workers accepted: 2290

Job description:
YouTube: Create an Account
1. **Go to www.youtube.com**
2. **Create a new account**
3. **Verify your email**
4. **Login to activate the account**

Required proof that task was finished?
1. **YouTube username or email**
2. **YouTube password**

Your task will NOT be rated satisfied if your YouTube account requests phone verification.

Campaign example 10: Youtube Account Creation

The client in this case has acquired 2,290 YouTube accounts for a mere gross $ 251.9. Example 2 from the introduction is a similar case, but for Gmail. The dangers posed by such campaigns are obvious. Besides promoting products, ideas and agendas, a cohort of 2,300 YouTube users can disrupt any smaller community on the platform and the use of fake accounts could facilitate the account owner to commit fraud or otherwise abuse the system.

What makes these mass account acquisitions very dangerous is that they are not easy to detect. Methods that are able to detect fake accounts typically only work if they are all created by the same person (Xiao u. a. 2015), but cannot be expected to work in this case as these accounts are created by real people. A landmark study by Gurajala u. a. (2015) involving the analysis of 62 million Twitter public user profiles relied on statistics about update frequency, reused profile pictures, and account creation days. Unfortunately, these factors can all be made to look genuine; for instance, the freelancers can be instructed to use profile pictures that are not reused; or possibly profile pictures themselves are acquired via microworkers (see example 13), and the creation times can be spread out with the help of "throttling"—a feature of the platform that allows only a certain number of tasks to be completed in a unit of time. Sometimes clients give instructions that enable the detection of such accounts, e.g., by requiring the freelancers to use the very same password. Also, it is probable that after handing over an account, the geolocation of the usage of that account is changed permanently, and so never again reflects the country of creation, which could be a factor in detection.

Not all signup campaigns seem to require account handover. For instance, the top four signup campaigns in terms of task numbers required a signup to two different website traffic providers and a polling site; involving 21,388, 19,952, 10,888, and 7,648 individual signups. Related to these campaigns was the biggest (99,999 workers) and cheapest (paying $0.0) campaign observed, categorized as testing (Z), as technically it was a website spellcheck; its details are given in the following example:

Title: Qualification Test: Find the Misspelled Word
Payment: 0.0
Number of workers accepted: 99 999

This is a qualification for a future website test which will pay $7.50 for about 11 minutes of work. This qualification test is to find workers with a great eye for detail.

1. **Go to REDACTED URL**
2. **Find the misspelled word.**
   Hint: it is near the bottom Required proof that task was finished?

1. **The wrongly spelled word in its wrongly spelled form**

Campaign example 11: A recruitment campaign involving a small test to pass

All five sites (the aforementioned four traffic providers and the one in Example 11) were categorized as "Other" on the platform, and were not commonly featured. The scale of these campaigns explain why the category Other is so prominent in the platform aggregation in Table 4. Also, all the sites are basically recruiting microworkers for their own platform. The nature of tasks to be done there seems to be are traffic generation (visiting sites), participation in paid market research by answering surveys, etc.

Among the next five in the top 10 signup campaigns (places 6-10 ranked by the number of tasks on offer) was example 10, another account creation and handover involving 2,190 accounts to a site redacted with the rotator technique, plus three jobs requiring the signup of several thousands of users to various sites for unknown reasons.

6.6 Other interesting campaigns

This section covers several interesting campaigns that cannot easily be categorized in the other categories, many of them one-of-a-kind campaigns, and some of them seem rather strange and unexplained.

There was one campaign that requires the users to solve captchas. This is obviously to bypass a captcha-protected signup page. We can hypothesize that this is part of a human-in-the-loop automated account creator system.

There were several research campaigns observed. These are transparent and benign: the university or the research group is clearly present, there is usually a document attached as a brief for the research. The topic seems to be social psychology or web usability and ergonomy. The users are asked their gender and then made to do face expression recognition; evaluate risks; try out different webpage workflows, etc.

There is one observed snapchat promoter recruitment campaign (30 tasks x USD 0.5), see the following:

Objectives: I'm looking for cute girls who snap for marketing promotions (bonus possible).
Important: You must actually snap video and not just upload pictures from fake profiles.
1. Provide your Snapchat Username for proof (I will add you as a friend)

Campaign example 12: Snapchat recruitment

Some campaigns seem to be building stock photos, like example 13 below (1000 x $ 0.11). Another project required photos of windshields. Yet another project asked for a selfie of the freelancer, and the consent to use it, but only from those who had no beard.

Important: You must agree to allow us to use your photo for promotional purposes in order to complete this task.
1. Take a well lit, clear photo of an office. Important: Make sure no people are in the photo.
Notes:- I need a clear photo of the office showing computers, desks, etc.- Photo should not be a photo from the Internet, we search for all photos on the Internet before we approve

Campaign example 13: Acquiring photos

Finally, for the following campaign there is just no explanation:
(30 x $1.75) Write 12 lines of lyrics for an Anthem based on your own individual traditions and struggles (Make it relevant to your life today) (...)

Campaign example 14: An unexplained campaign

7 Techniques employed in campaigns

As explained in the Limitations section, there are a number of invite-only campaigns on the site, called “hire groups”. These allow a client to select the freelancers, as contrasted to public campaigns that are open to anyone to participate. Also, this allows a per-employee task customization by providing a spreadsheet of input variable values. While also being feature rich, hire group campaigns are usually hidden from the public view.

Rotators are another way of per-employee customization and also allows hiding the content of the campaign from public display.

Title: Forum: Sign up + Post + Screenshot Payment: 0.14
Number of workers accepted: 96

1. **Go to this link: bit.ly/<REDACTED URL ENDING1>**
2. **Search for blogs from this search link**
3. **Find blog, website or forum you can post a comment on**
4. **Go to this link: bit.ly/<REDACTED URL ENDING2> and then copy comment from this page and post this comment in the website blog or forum**

Required proof that task was finished?

1. **Your Forum Username**
2. **URL of the comment**
3. **Screenshot of posting**

Campaign example 15: The rotator technique

This technique allows the employer to customize the task per-employer without a hire group and to remove the instructions after the campaign is done without leaving a trace. Except for those freelancers who participated in the campaign, there is no way of knowing what sites, search keywords, or comments were involved in the job. The category “Redacted” among the platforms refer to this technique and not to data anonymization employed by the author in the examples in this article. Of course, there is no way of knowing if the client’s intention was just to rely on task customization, or to hide the campaign content, or both.

As explained at the Smartphone apps section, there might be ways for dressing up promotion campaigns as testing campaigns, by asking a couple of hundred users to install the app and then leave it there. Also, there might be search and engage campaigns masquerading as data collection and competition monitoring. In the case of some Amazon- and eBay-related campaigns, the freelancers are directed to search for different products, then to select from a given set of results, and then to collect prices, data, specifications, and to finally submit these as job proof. What makes these suspicious is that for an honest information campaign, it seems to be overly redundant to collect the same information many hundreds of times by many hundreds of microworkers. In reality, the point of these campaigns could be to make the microworker search and engage and then to spend time on the visited page while counting reviews and collecting information (the algorithm of a search engine might take the duration spent on a result page into account when adjusting itself), and then the accomplishment of the job can be conveniently verified by the client by looking at the collected data. Of course, these are just hypotheses for which there is no way to verify them.

Figure 1 summarizes the promotional methods observed, together with the supporting techniques featured in various gray promotional campaigns:

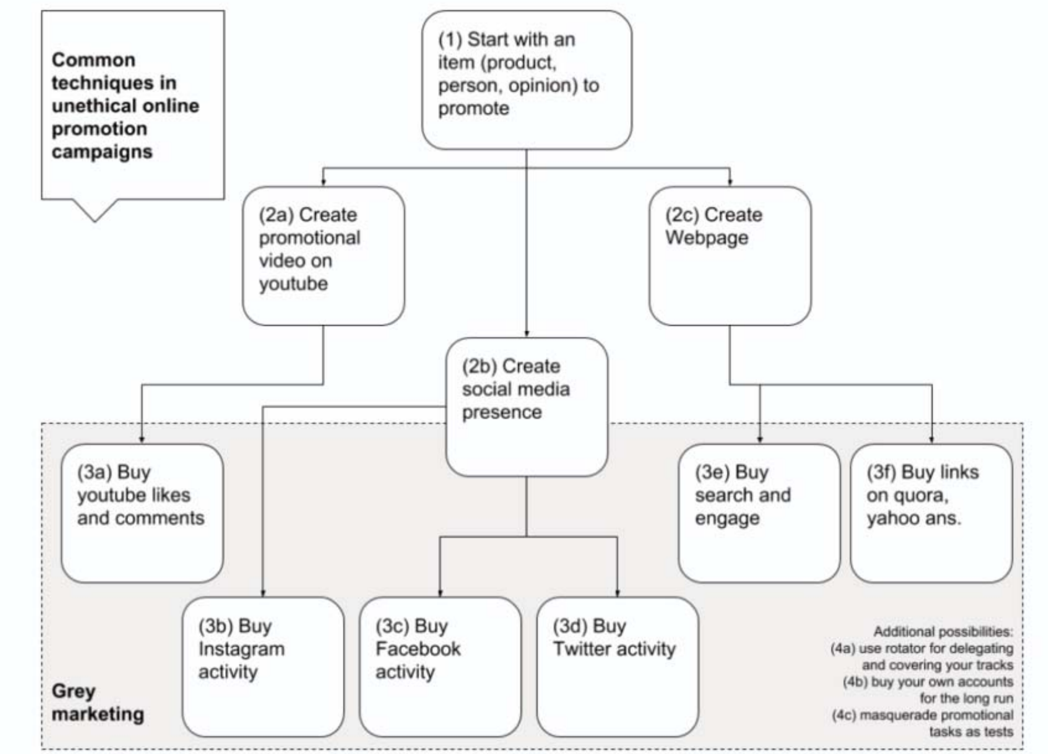

Common techniques in unethical online promotion campaigns
(1) Start with an item (product, person, opinion) to promote
(2a) Create promotional video on youtube
(2b) Create social media presence
(2c) Create Webpage
(3a) Buy youtube likes and comments
(3b) Buy Instagram activity
(3c) Buy Facebook activity
(3d) Buy Twitter activity
(3e) Buy search and engage
(3f) Buy links on quora, yahoo ans.
Grey marketing
Additional possibilities:
(4a) use rotator for delegating and covering your tracks
(4b) buy your own accounts for the long run
(4c) masquerade promotional tasks as tests

Figure 1: Elements of unethical online promotion campaigns.

8. Conclusion

This article provides insights into the black market of likes, upvotes, comments, retweets, votes in contests, and search engine manipulation. The subject of the investigation was microworkers.com, which is not a black market itself per se, but it has light regulation of its campaigns and so can be used by clients to participate in black-market activities. Also, it clearly is only one of several venues for running such campaigns. De Micheli und Stroppa (2013) have investigated several other players on the market (Fiverr, Seo- Clerks, Inter-Twitter, FanMeNow, LikedSocial, SocialPresence, SocializeUk, ViralMedia- Boost) and they have found that the market size is probably several millions of dollars, making the share microworkers.com a tiny fraction. Other sites even recruit on microwork- ers.com for similar microtasks. However, thanks to the fact that on microworkers.com the client has to orchestrate the campaigns itself, we can get an insight how the other players in the market, that sell complete like and follower packages, might be operating.

The nature of the microworkers.com campaigns was explained in the sections above. About their efficiency, we concluded that it probably varies. The main limitation is that it seems to be hard to purchase more than some tens of thousands of items. As explained in the section on Social media activity covering likes/upvotes (L), these numbers do not make a big difference when it comes to widely discussed political topics or celebrities, as in this area millions of L items are not uncommon. However, in smaller communities, with normally dozens or hundreds of L items, they can make a huge difference. This is the context in which the effects of a total of 207,811 L tasks can be assessed. For instance, Reddit, on which 77,104 upvotes were purchased, is a platform where a couple of hundred or thousand purchased upvotes can go quite far, especially in thematic sub-Reddits. It might be noted that a similar number of downvotes would be much more significant as there are normally much less of these items—but no downvoting/dislike campaigns were observed.

For comments, we have to assume that the big observed campaigns, reaching 60,000 YouTube comments must be effective as these are quite high numbers when it comes to comments. It is of course unclear what the overall effect of tens of thousands of comments is on the thinking of the targeted audience. But it is enough to provide an apparent majority on almost any platform.

For online voting and contests, it seems that all kinds except the biggest contests can be rigged by microworker campaigns.

There are two areas where the efficiency is especially hard to assess: search and engage and app testing. For search and engage, over half a million tasks were observed and we must assume that the efficiency of these jobs really depends on the popularity of the topic in question. Also, the effect of these campaigns is really hard to track. In a similar way, a hypothesis was provided on how app developers on Android or iPhone might by trying to build an initial user base of a couple of hundred installs. The most popular applications have tens of millions of user and even their alternatives often have tens or hundreds of thousands (this also indicates just how hard the entry must be to that market). A better understanding of the app market places would be necessary to understand the significance of a couple of hundred individual users.

Finally, the knowledge that several thousand YouTube, Gmail, Snapchat and other accounts have been created and their usernames and passwords handed over during the period of observation is very troubling. Those accounts might be effectively used for large scale gray promotion campaigns and could also pose a security threat at the same time.

Future work could involve participatory research as a freelancer on this or other platforms to reveal the experience of a microworker as well as to discover more about the invite-only/hire-only campaigns. In cases of comment copy-pasting, the source and nature of the comments could also be learned. Another area could be investigating the logic and goals behind the traffic generator sites and unions that similarly to microworkers' sites rely on the completion of menial tasks.

Acknowledgement

This work was supported by the MTA Bolyai scholarship and the ÚNKP-19-4 New National Excellence Program of the Ministry for Innovation and Technology.

References

Allcott, H. and M. Gentzkow, Social media and fake news in the 2016 election, *Journal of Economic Perspectives* 31. (2017) 2., 211–36.

Arthur, C., How low-paid workers at 'click farms' create appearance of online popularity, *The Guardian* . 2 August 2013.
https://www.theguardian.com/technology/2013/aug/02/click-farms-appearance- online-popularity

Buhrmester, M., T. Kwang and S. D. Gosling, Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?, *Perspectives on psychological science* 6 (2011) 1., 3–5.

Buttcher, S., C. L. Clarke and G. V. Cormack, *Information retrieval: Implementing and evaluating search engines*, Mit Press, 2016.

Caldwell, C. , Not being there, *The New York Times*. 12 August 2007.

Clark, J. , Google turning its lucrative web search over to AI machines, *Bloomberg Tecnology* 26. (2015).

Cook, D. M., B. Waugh, M. Abdipanah, O. Hashemi and S. A. Rahman, Twitter deception and influence: Issues of identity, slacktivism, and puppetry, *Journal of Information Warfare* 13. (2014) 1,: 58–IV.

Crone, D. L. and Williams, L. A. (2017). Crowdsourcing participants for psychological research in australia: A test of microworkers, *Australian Journal of Psychology* 69(1): 39–47.

De Micheli, C. and A. Stroppa, Twitter and the underground market, *11th Nexa Lunch Seminar,* Vol. 22. (2013), Politecnico di Torino.

Del Riego, A. , Digest comment-context for the net: A defense of the ftc's new blogging guidelines, *JOLT Digest, an online companion to the Harvard Journal of Law and Technology*. 2009.

Forrest, E. and Y. Cao, Opinions, recommendations and endorsements: The new regulatory framework for social media, *Journal of Business and Policy Research* 5. (2010), 2. 88–99.

Gardlo, B., M. Ries, T. Hobfeld and R. Schatz, Microworkers vs. facebook: The impact of crowdsourcing platform choice on experimental results, *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on,* QoMEX, 2012, pp. 35–36.

Gurajala, S., J. S. White, B. Hudson and J. N. Matthews, Fake twitter accounts: profile characteristics obtained using an activity-based pattern detection approach, *Proceedings of the 2015 International Conference on Social Media & Society,* Social Media Society, 2015, p. 9.

Hirth, M., T. Hobfeld and P. Tran-Gia, Anatomy of a crowdsourcing platformusing the example of microworkers. com, *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on,* IMIS, 2011, pp. 322329.

Howe, J. The rise of crowdsourcing, *Wired magazine*, 14. (2006).

Laperdrix, P., W. Rudametkin and B. Baudry, Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints, *Security and Privacy (SP), 2016 IEEE Symposium on,* IEEE, 2016, pp. 878–894.

Nguyen, N. , Microworkers crowdsourcing approach, challenges and solutions, *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, ACM, 2014, pp. 1–1.

Nicholas Confessore, R. H. Gabriel J.X. Dance and M. Hansen, The follower factory, *The New York Times* . 31 January 2018.

Paolacci, G., J. Chandler and P. G. Ipeirotis, Running experiments on amazon mechanical turk, *Judgment and Decision Making* 5. (2010) 5.

Rutz, O. J. and R. E. Bucklin, From generic to branded: A model of spillover in paid search advertising, *Journal of Marketing Research* 48. (2011), 1., 87–102.

Smith, D. (2018). Putin's chef, a troll farm and russia's plot to hijack us democracy, *The Guardian* . 17 February.
https://www.theguardian.com/us-news/2018/feb/17/putins-chef-a-troll-farm-and-russias-plot-to-hijack-us-democracy

Thackeray, R., B. L. Neiger, C. L. Hanson and J. F. McKenzie, Enhancing promotional strategies within social marketing programs: use of web 2.0 social media, *Health promotion practice* 9. (2008) 4., 338–343.

Xiao, C., D. M. Freeman and T. Hwa, Detecting clusters of fake accounts in online social networks, *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*, ACM, 2015, pp. 91–101.

Yang, S. and A. Ghose, Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence?, *Marketing Science* 29. 2010, 4., 602–623.

Zhang, Y., X. Li and T.-W. Wang, Identifying influencers in online social networks: The role of tie strength, *International Journal of Intelligent Information Technologies (IJIIT)* 9. 2013, 1., 1–20.

Author information

Mihaly Heder, Budapest University of Technology and Economics
https://orcid.org/0000-0002-9979-9101

Maintenance, function, and malfunction in technology

Alexandra Karakas

Abstract

This paper takes a new look at the concept of maintenance through the notion of function and malfunction. I propose that different maintenance strategies have contrasting philosophical approaches about the nature of technology. I claim that the main difference between reactive and predictive maintenance is that the latter holds an underlying deterministic approach.

This paper takes a new look at the concept of maintenance through the notion of function and malfunction. I propose that different maintenance strategies have contrasting philosophical approaches about the nature of technology. I claim that the main difference between reactive and predictive maintenance is that the latter holds an underlying deterministic approach. On the one hand, the assumption behind predictive maintenance is that we can predict how technology works and even progresses, and because of this, we can predict when will malfunction appear. On the other hand, reactive maintenance is based on unpredictability, and it is led by the appearance of failure. Thus, the central notion in both strategies is the concept of malfunction.

By reverse-engineering different predictive maintenance and reactive maintenance strategies, I point out a different notion of malfunction in these two strategies. In order to highlight the philosophical nature of maintenance theories, first I discuss how failure has been identified within philosophy of technology, and present how malfunction is related to and connect both the concept of function and maintenance. Although different maintenance descriptions can be highly technical, I claim that each method has an underlying philosophical idea behind it.

1. Introduction

Technological devices are usually described by their particular function. Thus, human-made objects all have specific descriptions of what they do and how they do it. However, occasionally technical artefacts cannot achieve the purpose they were designed for, and malfunction (including failure, and gradual degradation) appears as time goes by. In order to prevent malfunction, thus to keep artefacts functional, many different maintenance strategies exist within technology. Maintenance as a practice could mean restoration to its original functionality of the artefact, preserving, repairing, or even improvement of an object. Often maintenance tends to be studied as a more technical question, rather than a practice with philosophical implications. However, the paper argues that we cannot separate underlying theoretical assumptions about the nature of technology from practical maintenance measurements.

Central to every maintenance strategy is the notion of malfunction. Essentially, human do maintenance to avoid malfunction, thus to prevent gradual degradation, breakage, or sudden failure, and to keep artefacts well functioning for a more extended time. Broadly speaking[1] , there are two[2] significant types of maintenance[3] strategies: reactive maintenance and preventive maintenance . These refer to, on the one hand, different stages of the life-cycle of artefacts, and on the other hand, to diverse methods of preventing malfunction and possible degradation. These major strategies hold different premises regarding the relationship between technology and malfunction.

What reflects particular engineering and design choices in a technological object? Who is responsible for malfunction and failure in a technosystem? What happens when components of an object one by one get replaced? This paper takes a new look at the concept of maintenance through the notion of function and malfunction. I claim that the main difference between reactive and predictive maintenance is that the latter holds a deterministic approach. On the one hand, the assumption behind predictive maintenance is that we can predict how technology works and even progresses, and because of this, we can predict when will malfunction appear. On the other hand, reactive maintenance is about unpredictability, and it is led by the appearance of failure. By reverse-engineering different predictive maintenance and reactive maintenance strategies, I point out a different notion of malfunction in these two strategies.

The aim of the paper is to point out that the philosophical assumptions behind maintenance strategies can lead to practical consequences. Thus, theoretical assumptions and pragmatic manners cannot be separated in maintenance. Because of this, the paper relies on both philosophical works and more technical descriptions of technology. I analyse the relationship between various maintenance strategies and the notion of malfunction, to highlight the connections between function, malfunction, and maintenance. I propose that different maintenance strategies have contrasting philosophical approaches about the nature of technology[4]. Although different maintenance descriptions can be highly technical, I claim that each method has an underlying philosophical idea behind it. What they all share is some type of assumption about the notion of malfunction, which I argue is a crucial factor in technology.

Preventive maintenance strategies have a deterministic approach towards technological trajectories, while reactive maintenance holds a less predictable notion about the nature of technology. Reactive and preventive approaches are inherent to maintenance strategies as they serve as the philosophical basis of any maintenance method. The paper examines these two strategies in order to trace back the specific relations between practical outcomes and philosophical implications in major maintenance attitudes.

[1] There are many different ways authors differentiate maintenance strategies. For the purpose of the paper, I use Mobley's categorisation consequently.

[2] There are subcategories as well, but for the sake of brevity I examine the two major categories of maintenance, as these two represents the two different basic assumptions about the nature of technology.

[3] Keith R. Mobley, Maintenance Fundamentals (Burlington: Elsevier, 2004).

[4] The already existing deterministic-indeterministic distinction is applied to maintenance in order to examine the theoretical assumptions behind various maintenance strategies.

In order to highlight the philosophical nature of maintenance theories, first I discuss h ow failure has been identified within philosophy of technology, and present how malfunction is related to and connect both the concept of function and maintenance. In the next section, I discuss the general goals of maintenance, and then the underlying assumptions behind reactive maintenance strategies. In the last section, I examine the deterministic approach behind preventive maintenance strategies.

2. The notion of failure

The fundamental concept in any maintenance theory is the notion of failure and malfunction, because the primary goal of maintenance is to restore or to keep functionality, thus to make and keep the technological device be able to perform its intended function. Although the term *intended function* has some controversies[5] around it, but for the purpose of this paper, I use *required function*[6] as the intended, designed function of a technological device, thus a feature created by engineers and designers. In case of malfunction, only this *required* aspect of functional features is relevant. This also means that the required function is an accessible feature, and the user and the designer agree[7] about what the device supposed to do. While a lot of different functional features can be attached to one artefact beside its required function that could all work simultaneously, malfunction as a concept is rather settled in this sense. Thus, if a person would keep adding new usages now and then to an object, for instance, practical purposes or symbolical meanings, that would not cause any problem in the artefact's structure, but only add new layers to it. Functional features are extensible, but these do not add malfunctional features to the object at the same time.

Preston argues that artefacts tipically need to be 'maintained in order to continue to perform their functions effectively' [8]. Although, sometimes performing the required function results in need for maintenance (e.g., routine car maintenance), and in other cases the material of the object will degrade even without use (e.g., certain materials get rusty), and thus maintenance needs to be done. Hence, even normal functioning can result in malfunction and need of maintenance.

Although, at first sight, failure – especially from a user's perspective - seems to be a simple concept, but in reality both in philosophy of technology and in engineering failure

[5] For a new take on the concept of intended function see Public Artefacts, Intentions, and Norms in Maarten Franssen et al., Artefact Kinds: Ontology and the Human-Made World, Artefact Kinds: Ontology and the Human-Made World, 2014, 45–62, https://doi.org/10.1007/978-3-319-00801-1., and for another discussion of the problem see Beth Preston, A Philosophy of Material Culture. Action, Function, and Mind (New York: Routledge, 2013).

[6] Some authors, like Preston uses the terms proper and non-proper functions Preston, A Philosophy of Material Culture. Action, Function, and Mind.

[7] There are cases in which the user requires different purposes from the device than what the designer intended, and this causes malfunction and failure in the object, but for the sake of brevity in this article I do not go into details about these type of anomalies.

[8] Beth Preston, "Philosophical Theories of Artifact Function," in Philosophy of Science and Engineering Sciences, ed. Anthonie W.M. Meijers (Amsterdam: North Holland, Elsevier, 2009), 217.

as such is a wide-ranging phenomenon. Franssen gives a profound definition about malfunction that states that 'x is a malfunctioning K" expresses the normative fact that x has certain features f and that because of these features, if a person p wishes to achieve result of K-ing, then p has a reason not to use x for K-ing'[9]. Implementing maintenance into this formula would add a certain temporary element to the framework. Firstly, because a correct maintenance routine can prevent an artefact from being broken and stay a completely functioning object, e.g. car maintenance. Secondly, a maintenance routine in many cases can repair a failed artefact, or even improve the overall quality of a technological device, e.g. software maintenance. Thirdly, maintenance means in some cases transforming an object into a different artefact, e.g. old wooden furniture transformed into a new tool. In these cases, the person has a choice to use the object for an entirely different purpose.

These processes are led by the type of malfunction in each case. Indeed, there are clear cut cases in technology when something is considered a failure. Still, failure also can be partial, subtle[10] and unnoticed for an extended time. Birolini classifies [11] four causes of failure that maintenance has to account. The first is when an error is the symptom 'by which a failure is observed' [12]; the second is an intrinsic malfunction caused by human misuse. The third type accounts for different levels of failure, thus error of the device and failure at a higher level. The last kind is a physical process failing. A standard, less detailed differentiation[13] of failures is the ones caused by humans, and failures caused by design fail, aka equipment malfunction. As technology progresses, we tend to think that as designers are creating more and better equipment, structures, buildings and all kinds of technological devices, the chances of a disaster, failure, and accidents are getting lower and lower. However, the reality of our age cannot be further than this. Henry Petroski, author of *To Engineer is Human. The Role of Failure in Succesful Design* argues that in recent years there were many technological severe accidents, and he questions whether there is real technological progress at all [14].

The connection between failure and maintenance has been emphasised from a life-cycle perspective by many researchers. The lifespan and particular features of an artefact '(…) is now considered to be the designing engineer's concern, up till the final stages of the recycling and disposal of its components and materials, and the functional require-

[9] Maarten Franssen, "The Normativity of Artefacts," Studies in History and Philosophy of Science Part A 37, no. 1 (2006): 47, https://doi.org/10.1016/j.shpsa.2005.12.006.

[10] In design, there is a strange phenomenon between malfunction and users. We tend to adapt quickly to uncomfortable situations and barely operating objects. For instance, our car has several failed parts, and as time goes by, there is more and more error. However, we get used to the – among other things - malfunctional airconditioning, the failed window electronics, and so on. As soon as a stranger arrives in the car, she immediately notices all the failure in the vehicle.

[11] A Birolini, Reliability Engineering: Theory and Practice (Springer, 1999), 3–4.

[12] Birolini, 3.

[13] Indeed, there is also failure caused by nature, such as tornados and extreme weather conditions, but this aspect of the topics is skipped here for the sake of brevity.

[14] Henry Petroski, To Engineer Is Human: The Role of Failure in Successful Design (New York: Vintage Books, 1992), 2, https://doi.org/10.1086/354258.

ments of any device should reflect this.'[15] Thus, the responsibility of the engineer does not end with creating and selling a design but includes the afterlife of the object, including the maintenance instructions as well.

3. Maintenance types

According to the European Standard, a basic definition of maintenance is that it is a 'combination of all technical, administrative and managerial actions during the life-cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function'. [16] Malfunction refers to, on the one hand, retaining a technical artefact in a preferred state before any type of malfunction occurs, and on the other hand, recover the object after malfunction occurred. Required function refers to the function or combination of functions of an item which are considered necessary to provide a given service connected to the artefact. Thus, required functions specify what activities the object must be able to fulfil.

The main goals of maintenance are (1) the reduction of the chances of failure, (2) to give information about the object's ideal use, and (3) about the stage of the life-cycle of the artefact, (4) to minimise the amount of time spent with inactivity because of failure occurrence, (5) to be cost-effective, and (6) to make the best out of an artefact's performance. Namely, to have a technical device operating for as long as possible. Altough sometimes conflicting, these aims are always densely interwoven, especially in the case of sophisticated devices and technological systems. 'The complexity of a device will affect how difficult it will be to maintain or repair it, and ease of maintenance or low repair costs are often functional requirements.' [17]

Maintenance can be divided into many different[18] categories [19], for example, *online versus on site* maintenance. It can also be approached from a professional and a non-professional, thus from a user's perspective. In this paper, I use the two traditional broad maintenance categories: reactive and preventive, maintenance for the sake of brevity. These diverse types of maintenance strategies refer to different approaches towards artefacts, and also to practices aimed at different life stages of objects.

[15] Ibo Franssen, Maarten, Lokhorst, Gert-Jan and van de Poel, "'Philosophy of Technology', The Stanford Encyclopedia of Philosophy (Fall 2018 Edition)," 2018, https://plato.stanford.edu/entries/technology/.

[16] EN 13306:2010, "European Standard – Maintenance Terminology," 2010, 5.

[17] Franssen, Maarten, Lokhorst, Gert-Jan and van de Poel, "'Philosophy of Technology', The Stanford Encyclopedia of Philosophy (Fall 2018 Edition)."

[18] See for instance Neil Bloom, Reliability Centered Maintenance. Implementation Made Simple (McGraw-Hill Companies, 2006)., Luca Del Frate, "Failure of Engineering Artifacts: A Life Cycle Approach," Science and Engineering Ethics, 2013, 913–44, https://doi.org/10.1007/s11948-012-9360-0.

[19] Mobley, Maintenance Fundamentals.

3.1. Reactive Maintenance

Mobley calls the reactive approach run-to-failure management, [20] as the logic of it is just to react and try to solve the problem when disaster and failure appear. This method is the oldest of maintenance strategies. However, in practice, this is a no-maintenance-maintenance, as this approach does not require any planning beforehand, unlike the rest of the maintenance methods. Proven that reactive maintenance is the most expensive [21] of all the maintenance strategies, it runs counterintuitive that it is still widely used in technology. There are many problems with reactive maintenance. The first issue is the cost-benefit ratio: in these terms, reactive maintenance is surprisingly inadequate. 'The major expenses associated with this type of maintenance management are (1) high spare parts inventory cost, (2) high overtime labour costs, (3) high machine downtime, and (4) low product availability.' [22] Spare parts are crucial as reactive maintenance supposed to be prepared for different possible malfunctional occurrences. In the case of reactive maintenance, the type of failure determines how we treat a technological device and what kind of actions have to be done.

How can we reverse engineer reactive maintenance strategies? As opposed to preventive maintenance, reactive maintenance is driven by malfunction, meaning that reactive maintenance only happens once failure appeared. It implies a more chaotic, less deterministic approach towards technology for two reasons. The first is the intuitive premise that nobody would let failure happen if he knows that it will happen for sure. That is to say, if an engineer knows the chances of a malfunctional occurrence within a technological system, and the possibilities are quite high, it is more than likely that he would want to prevent any failure before it happens. Reactive maintenance exists only because in many cases, humans just cannot predict what will happen with a technological invention, or more precisely how technology will progress. The other reason behind reactive maintenance is as simple as it gets: the lack of detailed knowledge about technology, interfering throughout systems, or in some cases about the properties of particular materials. With reactive maintenance, the term can be divided into two different aspects of malfunctional occurrences: (1) failure caused by humans, and (2) failure caused by different actors (e.g. natural disasters).

The first type of malfunction can also be subcategorised: failure caused by the engineer (designer), and the ones caused by users[23]. They all essentially form risk management. However, there are cases when different levels of responsibility and failure are intertwined. This happened with one of history's most severe disasters, the Chernobyl Nuclear Power Plant accident. In that case, engineers were planning to do maintenance. Still, before that, they were running tests 'to study the possibility of utilisation of the mechanical energy of a turbogenerator after the cut-off of steam supply, in order to ensure

[20] Mobley, 2.

[21]Laura Swanson, "Linking Maintenance Strategies to Performance," International Journal of Production Economics 70, no. 3 (2001): 238, https://doi.org/10.1016/S0925-5273(00)00067-0.

[22] Mobley, Maintenance Fundamentals, 2.

[23] Another example of reactive maintenance is when the actual design of an artefact dictates how maintenance can be done, for instance with Apple products.

the power requirements in a case of a power failure' [24]. According to various studies, the cause of the Chernobyl disaster is varied: on the one hand, it caused by design mistakes of the reactor and misinformation about safety regulations, and on the other hand, mistakes made during the testing of the generators [25] by inadequately trained staff. In this case, test before even doing the real maintenance caused a disaster that requested quick reactive strategies from the team.

3.2. Preventive maintenance practices and their implications

The primary characteristic of preventive maintenance strategies[26] is that they are all time-driven and have a specific deterministic character. They need to be done at fixed intervals that are based on particular criteria, let it be daily, monthly, yearly maintenance, or any other fixed range. The point of preventive maintenance is to decrease the probability of error, failure, or any kind of malfunctional occurrence of a yet functioning item. The basic idea behind preventive maintenance is that there is a sometimes undetectable 'cause-and-effect relationship between scheduled maintenance and operating reliability' [27]. This assumption is based on two hypotheses: the first one is the intuitive belief that the older an object gets, the chances are higher of failure, and the second is that replacing older parts of an object would prevent failure.

Predictive maintenance is a specific type of preventive maintenance strategies. The most commonly used definition of predictive maintenance is the following: 'condition-based maintenance carried out following a forecast derived from repeated analysis or known characteristic and evaluation of the significant parameters of the degradation of the item.'[28] Predictive maintenance can be described as a means of improving systems and particular devices, and in general greater technosystems. Usually, predictive maintenance consists of different tools[29] that produces factual data on technology, and in many cases, predictive maintenance prevents sudden, unscheduled breakdowns of machines [30]. The underlying philosophical assumption behind this maintenance strategy is that it is possible to detect malfunction before it occurs. Thus, if objects are being analysed regularly, the chances of failure can be reduced. A well-established predictive maintenance

[24] Mikhail V. Malko, "The Chernobyl Reactor. Design Features and Reasons for Accident," 2016, 16–17, https://inis.iaea.org/search/search.aspx?orig_q=RN:48080457.

[25] Malko, 11.

[26] NASA also uses the so-called Reliability-Centered Maintenance that encompasses PM, Predictive Testing and Inspection, Proactive Maintenance and Repair at the same time in order to minimise the chances of malfunction. The discussion of this strategy is skipped here for the sake of brevity. For a detailed description, see NASA Reliability-Centered Maintenance Guide for Facilities and Collateral Equipment.

[27] "NASA Reliability-Centered Maintenance Guide for Facilities and Collateral Equipment," 2008, 2–1.

[28] 13306:2010, "European Standard – Maintenance Terminology," 12.

[29] Mobley claims that there are five techniques of predictive maintenance: vibration monitoring, process parameter monitoring, thermography, tribology, and visual inspection. The analysis of these more technical questions is skipped in this paper.

[30] Mobley, Maintenance Fundamentals, 5.

strategy 'utilises the most cost-effective techniques in a combination to obtain the condition of critical equipment.' [31] The data gathered through the different monitoring activities provide sufficient information about the state of the device.

Contrary to reactive maintenance practices, the underlying assumption behind preventive maintenance strategies is a deterministic approach. This is based on the idea that technology as a system and the advancement of particular devices hold trajectories, and humans can know and manipulate these paths for the better. However, technology is continuously shaped by influences from different actors, such as users, designers, and economic factors. The effects of these different spheres are profoundly interwoven. Broadly speaking, in the field of philosophy of technology[32] researchers divided technological changes into two different categories: technology-push, and demand-pull approach. The first emphasises social forces as crucial factors in technological change, while the latter defines technology as an autonomous entity. Although, Dosi claimed that these categories are inadequate for understanding and explaining technology, and he argued for a different approach based on scientific paradigms [33].

Inspired by Thomas Kuhn's *The Structure of Scientific Revolutions*. Dosi explained that technology has particular trajectories, similar to "normal science" period in Kuhn's theory and that these trajectories also define in which direction science progresses. Dosi described technological trajectories as a 'cluster of possible technological directions whose outer boundaries are defined by the nature of the paradigm itself' [34], meaning that there can be stronger and weaker trajectories as well as complementarities between trajectories. Dosi's framework has strong instrumental values in contextualising technological problems. Predictive maintenance strategies often imply a determined trajectory to technological progress. In many cases, they do not count with random factors such as in the case of the Chernobyl Nuclear Power Plant accident. Nearby technological trajectories can diverge or slowly move away from each other because of several reasons. Are trajectories a real feature of technology, or they only have instrumental value?

Summary

Even though malfunction is a crucial feature of technology, its role and status in particular devices and higher systems are still unclear. What causes is that in the majority of technological failures, engineers and average users can only retrospectively investigate the nature

3

[1] Gustav Fredriksson and Hanna Larsson, "An Analysis of Maintenance Strategies and Development of a Model for Strategy Formulation – A Case Study" (CHALMERS UNIVERSITY OF TECHNOLOGY, 2012), 31.

[32] For a review and bibliometric analysis of the role of demand in technology and innovation see Technology Push and Demand Pull Perspectives in Innovation Studies: Current Findings and Future Research Directions Giada Di Stefano, Alfonso Gambardella, and Gianmario Verona, "Technology Push and Demand Pull Perspectives in Innovation Studies: Current Findings and Future Research Directions," Research Policy 41, no. 8 (2012): 1283–95, https://doi.org/10.1016/j.respol .2012.03.021.

[33] Giovanni Dosi, "Technological Paradigms and Technological Trajectories," Research Policy 11, no. 3 (1982): 147–62, https://doi.org/10.1057/978-1-349-94848-2_733-1.

[34] Dosi, 154.

of certain malfunctional occurrences. With new technologies, new failures appear as well, and in many cases detecting the exact cause of failure is a tricky thing. Among many other technological error[35], software engineering is one of the most common sources of failure nowadays. The goal of the paper is to point out that different philosophical assumptions behind maintenance strategies can have practical consequences as well. The effects of the deterministic nature of preventive maintenance strategies are that it can easily lead to simplistic notions of technology. Because of this, it can overlook serious problems concerning artefacts.

In this paper, I claim that the main difference between reactive and predictive maintenance is that the latter hold a deterministic approach. On the one hand, the assumption behind predictive maintenance is that we can predict how technology works and even progresses. Because of this, we can predict when will malfunction appear. On the other hand, reactive maintenance is about unpredictability, and it is led by the appearance of malfunction. By reverse-engineering different predictive maintenance and reactive maintenance strategies, I point out a different notion of malfunction in these two strategies. Identifying theoretical assumptions behind maintenance strategies can help reduce the chances of failure, and also to lay the foundations of a comprehensive understanding of the notion of malfunction.

Bibliography

13306:2010, EN. “European Standard – Maintenance Terminology,” 2010.

Birolini, A. *Reliability Engineering: Theory and Practice*. Springer, 1999.

Bloom, Neil. *Reliability Centered Maintenance. Implementation Made Simple*. McGraw-Hill Companies, 2006.

Charette, Robert N. “The Biggest IT Failures of 2018,” 2018.

Dosi, Giovanni. “Technological Paradigms and Technological Trajectories.” *Research Policy* 11, no. 3 (1982): 147–62. https://doi.org/10.1057/978-1-349-94848-2_733-1.

Franssen, Maarten, Lokhorst, Gert-Jan and van de Poel, Ibo. “‘Philosophy of Technology’, The Stanford Encyclopedia of Philosophy (Fall 2018 Edition),” 2018. https://plato.stanford.edu/entries/technology/.

Franssen, Maarten. “The Normativity of Artefacts.” *Studies in History and Philosophy of Science Part A* 37, no. 1 (2006): 42–57. https://doi.org/10.1016/j.shpsa.2005.12.006.

Franssen, Maarten, Peter Kroes, Thomas A.C. Reydon, and Pieter E. Vermaas. *Artefact Kinds: Ontology and the Human-Made World. Artefact Kinds: Ontology and the Human-Made World*, 2014. https://doi.org/10.1007/978-3-319-00801-1.

Frate, Luca Del. “Failure of Engineering Artifacts : A Life Cycle Approach.” *Science and Engineering Ethics*, 2013, 913–44. https://doi.org/10.1007/s11948-012-9360-0.

Fredriksson, Gustav, and Hanna Larsson. “An Analysis of Maintenance Strategies and Development of a Model for Strategy Formulation – A Case Study.” CHALMERS UNIVERSITY OF TECHNOLOGY, 2012.

Malko, Mikhail V. “The Chernobyl Reactor. Design Features and Reasons for Accident,” 2016. https://inis.iaea.org/search/search.aspx?orig_q=RN:48080457.

[35] For a selection of failures in technology in 2018, see The Biggest IT Failures of 2018 Robert N. Charette, “The Biggest IT Failures of 2018,” 2018.

Mobley, Keith R. *Maintenance Fundamentals*. Burlington: Elsevier, 2004.

"NASA Reliability-Centered Maintenance Guide for Facilities and Collateral Equipment," 2008.

Petroski, Henry. *To Engineer Is Human: The Role of Failure in Successful Design*. New York: Vintage Books, 1992. https://doi.org/10.1086/354258.

Preston, Beth. *A Philosophy of Material Culture. Action, Function, and Mind*. New York: Routledge, 2013. .

"Philosophical Theories of Artifact Function." In *Philosophy of Science and Engineering Sciences*, edited by Anthonie W.M. Meijers. Amsterdam: North Holland, Elsevier, 2009.

Stefano, Giada Di, Alfonso Gambardella, and Gianmario Verona. "Technology Push and Demand Pull Perspectives in Innovation Studies: Current Findings and Future Research Directions." *Research Policy* 41, no. 8 (2012): 1283–95. https://doi.org/10.1016/j.respol.2012.03.021.

Swanson, Laura. "Linking Maintenance Strategies to Performance." *International Journal of Production Economics* 70, no. 3 (2001): 237–44. https://doi.org/10.1016/S0925-5273(00)00067-0.

Author information
Alexandra Karakas, Eötvös Lóránd University
http://elte.academia.edu/AlexandraKarakas/CurriculumVitae

Responsibility in biomedical engineering education: a comparative study of curriculum in India, Russia and the USA

Aleksandra Kazakova

Abstract

This article adds a comparative perspective to the ongoing debate on the ways of integrating the principle of responsibility into engineering education. While the need for increasing social and environmental awareness is expressed both by the professional and educational communities, the concrete measures for restructuring the curricula raise methodological, epistemological and pedagogical questions. The study links the normative debate to the state of the art in biomedical engineering curriculum in three different educational systems. The content analysis shows that in contrast to the commonly declared objective of formation of a responsible engineer, the range of disciplines and subjects that could contribute to its achievement is underrepresented in the undergraduate programs in terms of the workload. Despite the differences in the national standards, the level of institutionalization remains equally low.

Keywords: Engineering education, engineering ethics, educational standards, ABET criteria

Introduction

The need for broader engineering education is expressed by the scientific and educational communities, professional associations and accreditation bodies. At the beginning of the century, a wave of strategic documents was setting the goals for reforms at the various levels and reconsidering the principles of professionalism with regard to the visions and expectations for sociotechnical future, e. g. “Educating the Engineer of 2020” by the National Academy of Engineering (2005), “Educating Engineers for the 21st Century” by the Royal Academy of Engineering (2007), “Engineering: Issues, Challenges, and Opportunities for Development” by UNESCO (2010), just to name a few.

The normative aspects of developing the engineering curriculum were conceptualized as “a new occupational ideal of Bildung for engineers” (Christensen, Meganck and Delahousse 2007, 13) which would guide the formation of the responsible “new engineer” (Beder 1998), or even give rise to the “post-engineering” culture (Mitcham 2009). The common idea in this debate is that engineering education needs to overcome the narrow problem-solving approach by recognition of both its societal context and impact, and thus to increase reflexivity of the actors producing and applying technologies.

This work outlines the current level of integrating the concept of professional responsibility in biomedical engineering curriculum in different national contexts. The first part is devoted to an overview of the methodological debate on the ways of teaching “responsible engineering”. It is shown that there is no final consensus on the optimal strategy of designing the engineering curricula, but the courses in engineering ethics alone, despite of their relatively high institutionalization, are increasingly regarded as insufficient for this purpose. Rather, a range of courses in social and environmental sciences and humanities are considered "responsible" for responsibility of the future engineers, and there are also calls for changes in professional training on the whole, such as "ethics across the curriculum". In the second part, the qualitative analysis of the actual educational programs in biomedical engineering is undertaken to assess the extent to which responsibility-related disciplines and subjects are represented in the curriculum. This work follows the national curriculum studies on the sets of mandatory disciplines (Stephan 1999) and on the structure of credit hours (Prasad et al. 2018), contributing the international comparison in one particular field of the engineering education. The analysis of the officially published programs certainly gives a preliminary picture of the institutionalization process, while the content and quality of educational practices require in-depth survey and assessment of the stakeholders.

Teaching "responsible engineering": the methodological debate

An extensive conceptual and empirical research on the different components of engineering curriculum has been made internationally. The methodological debate about the structure of engineering education in general and the responsibility-related subjects, in particular, is driven by the necessity to find the best possible set under existing limitations of time and resources in the context of growing specialization and competition on the global educational market. This stimulates discussion of the comparative advantages of different disciplines and the arguments legitimating the various ways of socio-humanitarian "intervention" into technical education.

Engineering ethics has been in the centre of this debate, especially since 2000, when Accreditation Board for Engineering and Technology (ABET) explicitly included ethics-related outcomes into its accreditation criteria, though without giving concrete recommendations on the forms and content of teaching (Mitcham 2009). Prior to the implementation of ABET criteria, a catalog-based survey of the engineering programs in the USA was undertaken, which showed "the relative invisibility of ethics-related instruction in present course requirements" (Stephan 1999, 459), since less than one quarter of institutions required all their students to take any course addressing this topic.

Almost 10 years later, Bucciarelli (2008, 148) argued that "the way ABET's recommendation for the study of ethics has been implemented within engineering programs falls far short of the mark", since the wide-spread methodology of case studies and teaching codes in the engineering ethics courses reproduced the abstract problem-solving approach, reducing the complexities of the socially contextualized engineering practice and communication processes to individual decision-making. This narrow didactic approach along with the problems of restructuring the curriculum led to the situation in which the engineering students "seldom take, and are certainly not required to take, courses dealing with the historical and social character of public safety, public health, or societal welfare" (Mitcham 2009, 36), which they were considered responsible for. As a result, focus on teaching micro-ethics which emphasizes individual responsibility was increasingly criticized in the last two decades with repeated calls "to consider questions of macro-ethics related to institutional organizations and public policy" (Mitcham and Englehartd 2016, 1739).

Some researchers went further arguing for diversification of the socio-humanitarian courses in the curriculum. For instance, Conlon (2008) argued for integration of social sciences into engineering education to make it able to reflect on the social nature of the technical problems and solutions and their relation to social conflict, inequality and power. Teaching ethics alone, he claimed, is thus insufficient without the focus on "the social structure and the way it both enables and constrains socially responsible conduct" (Conlon 2008, 151). At the same time, he warned against the utilitarian approach to the socio-humanitarian component of the curriculum, which may replace the goal of developing responsibility with that of increasing employability through development of non-technical competencies, or the so called "soft skills" (such as communications, project management, leadership and teamwork).

In contrast to teaching a general course in social sciences, some scholars argued for more focused courses in STS. For example, Pinch elaborated the course which combined the key STS concepts, "based upon deep sociological ideas which often extend well beyond the boundaries of science, technology, and medicine" with the relevant case studies, thus showing "how these concepts can be used in many different contexts and historical periods" (Pinch 2008, 104-5). With its multidisciplinary nature, STS course would touch upon not only sociological, but also philosophical, anthropological, historical and political perspectives on science and technology. Apart from that, he emphasized that many STS researchers have scientific or engineering background. This may facilitate communication with engineering audience.

Downey insisted that engineering studies need to overcome its marginal position in the curriculum and "open up engineering formation" itself, hence not only providing the critical analysis of the engineering activity in society, but also reflecting on the educational process as such "to make visible the value dimensions of engineering pedagogies inside and outside of classrooms" (Downey 2015, 218). He claims that this should go beyond "contextualizing" of engineering and overcome the very distinction between technical and social. This intervention is a part of wider "scalable scholarship" in engineering studies, which must "contest the dominant epistemological contents of engineering practices" (Downey 2009, 55).

Probably, even more radical steps were suggested by Bucciarelli and Drew (2015) who elaborated a program for Bachelor of Arts in Liberal Studies in Engineering, which would not only synthesize education in humanities, social sciences and engineering, but also make it more inclusive by reducing the barriers, explicating values behind engineering activity and assigning meanings to it. Even if this Renaissance-like project will not be fully realized on the contemporary educational market, the very idea of interdisciplinarity and change of the teaching practices in the core professional (technical) disciplines is not so odd. It is close to the ideas of "ethics across the curriculum" and "sociotechnical design", which do not require such a profound institutional transformation.

The "ethics across the curriculum" approach, despite of its popularity and visible organizational efforts since 1980s, is still not clearly conceptualized and has received diverse interpretations and applications in different fields of education. Apart from other pedagogical innovations, one of the objectives for the faculties was to "increase the inclusion of ethics in courses taught and integrate the discussion of ethical issues with standard subject matter" (Mitcham and Englehartd 2016, 1745). If applied systematically, this diffusive strategy could stimulate ethical inquiry at the level of everyday learning routines and thus change the culture of engineering education from within.

In line with this approach, Date and Chadrasekharan (2017) suggest that developing commitment to sustainability requires more than just didactic statement of principles. It requires change of teaching at epistemological level, transforming the problem-solving approach into "solving for pattern". They argue that "shift towards sustainability engineering requires illustrating successful design practices that embed sustainability values, particularly designs that move away from the current focus on input–output efficiency, towards eco-social and socio-technical approaches to design" (Date and Chadrasekharan 2017, 12).

All the touched upon epistemological and ethical issues, as well as questions of engineering culture and practices, and the wider context of sociotechnical transformations are discussed in philosophy of science and technology, which also claims for its place in engineering curriculum. It was characterized by Christensen and Ernø-Kjølhede (2008, 563) as a "Socratic element of professional self-reflection in engineering education". However, their empirical study of expectations of an engineering faculty from implementation of this course shows that the attitude to philosophy reproduces rather instrumentalist approach to education in general. Still, the authors hope that further integration of philosophical courses may compensate this "lack of metadiscourse", or "metalevel perspective" in engineering education and engineering as such.

The overview of the debate among the scholars and educationalists permits the following conclusion: there is no royal road to responsible engineering. Apart from the obvious contribution of engineering ethics, alternative ways of integrating the responsibility agenda into the curriculum have been suggested. The wide range of disciplines and interdisciplinary courses in social and environmental sciences and humanities can be regarded as responsibility-related. Apart from this, such approaches as "ethics across the curriculum" and "eco-social design" aim to problematize social and environmental issues in professional training. At the same time, the socio-humanitarian component may be reduced to the soft skills training courses and thus instrumentalized as employability-oriented. For preliminary assessment of the level of responsibility-oriented instruction it is therefore important to combine the study of curriculum structure with the content analysis of the syllabus.

Responsibility in curriculum: design of the study

Following the methodological debate described above, the study was designed to trace the various ways of integrating the principles of social and environmental responsibility into the curriculum. It is assumed that this may be achieved by systematically addressing the problems of ethics, sustainability, societal or environmental risks and safety in general sense (not reduced to safe working conditions) in different disciplines.

The content analysis of 48 baccalaureate programs in bioengineering, biomedical engineering or similarly named engineering programs in three countries (the USA, Russia and India) was undertaken. The special interest in this field of engineering education stemmed from an idea that biomedical engineering has the longest tradition of ethical inquiry, inherited from the medical education. When compared to other fields of engineering, it implies the most obvious "human-machine" interaction, which requires special concern for safety of a user/patient. With that in mind, it can be assumed that biomedical specialization is especially sensitive with regard to the problems of responsibility.

Only educational programs accredited by the national or international bodies (ABET, National Board of Accreditation in India, Russian Association of Engineering Education) were examined, assuming their legitimacy in the eyes of professional and educational communities. The publication of the program documents, including their missions, expected educational outcomes, curriculum and syllabus, on the official websites of the institutions are regarded as both the representation of the educational strategies and the minimum level of institutionalization of the responsibility-oriented instruction. The latest version of officially published educational programs was analyzed (academic year 2018/2019).

Three questions were posed for analysis:

1. Is development of social and / or environmental responsibility explicitly declared among the goals of the educational program?

2. What is the share of the workload for mandatory courses specialized in the problems of social and environmental responsibility (ethics, sustainability, societal / environmental risks / safety)?

3. Are the relevant topics discussed in non-specialized (introductory or other professional) courses?

To answer the first question, a qualitative analysis of missions, objectives, outcomes and competences was made. For the second and third questions, the catalogues of courses, their abstracts and schedules were examined.

On the basis of the previous methodological debate, the category of "specialized" courses was defined broadly, including courses in ethics ("Professional Ethics", "Engineering Ethics", "Bioethics") as well as in philosophy and sociology, STS, environmental studies, safety and risks, technology assessment or integrated courses, combining the listed above. Regardless of their belonging to socio-humanitarian component, the "soft skills" courses (writing, communication, rhetoric, time management, language courses, entrepreneurship), as well as the courses in economics, law, IPR and management were not taken into account as related to employability and liability rather than responsibility (see the discussion above). The courses in national history, political system and diversity were also excluded, being not specific for professional activity. For the non-specialized courses the syllabus descriptions were analyzed in search of the relevant topics (keywords: "responsibility", "sustainability", "ethics", "risk", "safety" and the cognates).

Results

Most of the programs in the US contain a standard list of educational objectives and learning outcomes based on the official ABET criteria of previous years (ABET, 2016; ABET, 2017), such as "an ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability", or "the broad education necessary to understand the impact of engineering solutions in a global, economic, environmental, and

The USA

Reference to social/environmental responsibility	**Number of programs**	**% of total number**	**Average working load (% of total program credits)**
Declared objective/outcome	21	100%	
At least one mandatory specialized course	11	52%	2.2%
* two or more specialized courses	0	0%	0%
Part of an introductory course	5	24%	N/a
No mandatory (specialized or introductory) course	7	33%	N/a
References in the other professional courses	7	33%	N/a
Total (programs)	**21**		

Table 1. Reference to social / environmental responsibility in the educational programs in biomedical engineering (the USA)

societal context", etc. Some of the programs reformulated the ABET list in more concise manner and with minor variations. However, all the examined programs explicitly declared responsible professional activity as priority.

11 educational programs included at least one mandatory specialized course addressing the topics in question, even though their share of workload was relatively small, for example: *Professional Responsibilities of Engineers* (3 of 186 Credits), *Safety and Ethics for Research* (1 of 182 Units), *Biomedical Engineering in the Real World* (1 of 129 Credits), *Biomedical and Bioengineering Ethics* (1 of 120 Credits, lectures only).

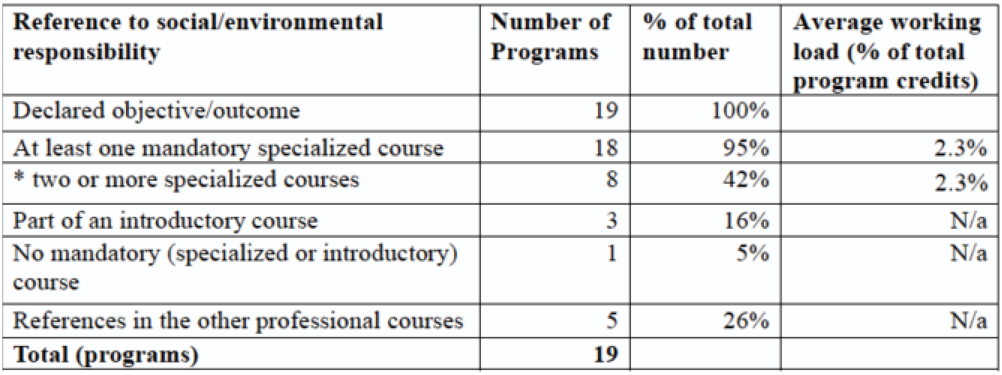
5 programs offered short introductory courses for the first or second year, which are sometimes department- or college-wide (that is, not specific for biomedical engineering), but explicitly address the problems of professional responsibility according to their synopsis: *Professional Development in Engineering* (2 of 126 Credits; the annotation says: "… about one-third of the semester is dedicated to professionalism and ethics"), *Professional Communication for Engineers* (1 of 133 Credits), *Engineering Success for First-Year Students* (1 of 128 Credits), *Engineering Disciplines and Skills* (2 of 128 Credits), etc. However, with one exception mentioned above, it is not clear what share of the working load is actually devoted to this topic within the course.

"Ethics-across-curriculum" approach. 7 institutions have made visible efforts to address the problems of responsible engineering in the professional courses. However, often these are the practical courses of the last years, i.e. the students are required to make an assessment of their individual or group projects without previous systematic training.

Non-mandatory / elective courses. The overall share of credits in Humanities and Social Sciences in the sample reached 13% of the programs' total. However, this was achieved by the variety of elective courses. In order to estimate the comparative popularity of the electives on professional responsibility, additional attendance data is needed.

According to the data, 7 of 21 programs did not include any mandatory (specialized or introductory) courses on professional responsibility, except for possibly chosen electives and occasional mentioning in professional courses (3 programs).

India



Reference to social/environmental responsibility	Number of Programs	% of total number	Average working load (% of total program credits)
Declared objective/outcome	19	100%	
At least one mandatory specialized course	18	95%	2.3%
* two or more specialized courses	8	42%	2.3%
Part of an introductory course	3	16%	N/a
No mandatory (specialized or introductory) course	1	5%	N/a
References in the other professional courses	5	26%	N/a
Total (programs)	**19**		

Table 2. Reference to social / environmental responsibility in the educational programs in biomedical engineering (India)

19 programs were examined, 18 of which are accredited by the National Board of Accreditation, and one by the ABET. Regardless of being accredited by ABET or by the national accreditation body, all the Indian programs contain similar lists of Program Educational Objectives (PEOs) and Program Outcomes (POs) which are close to ABET criteria. Typically, it includes understanding of "ethical and professional responsibility" as well as of "impact of engineering solutions in a global, economic, environmental, and societal context" and "realistic constraints such as economic, environmental, social, political, ethical, health care and safety, manufacturability, and sustainability" with minor modifications.

A significant share of the programs (8 out of 19) combines both social and environmental aspects of responsibility, offering more than one mandatory specialized course, for example: *Professional Ethics and Human Values* and *Introduction to Environmental Science* (6 of 173 Credits), *Introduction to Society and Culture* and *Environment and Safety Engineering* (6 of 215 Credits), etc. However, their share in the overall workload is close to that in the US (2.3%). One of the programs has mandatory, but no-credit courses in both areas. As for the rest, 10 programs contained only one specialized course, and with one exception, most of them were focused on environmental responsibility or sustainability.

"Ethics-across-curriculum" approach. Only in 5 programs explicit and systematical reference to the problems of responsibility was found in the professional courses, such as *Stem Cells and Healthcare* or *Decision Support Systems*.

16 programs contained relevant elective courses, such as *Environmental Impact Assessment*, *Human Factors in Engineering*, *Engineering Law and Ethics*, *Engineering and Society*, *Green and Sustainable Development*, but no data is available to estimate demand for them.

Russia

Reference to social/environmental responsibility	**Number of Programs**	**% of total number**	**Average working load (% of total program credits)**
Declared objective/outcome	8	100%	
At least one mandatory specialized course	8	100%	3.4%
* two or more specialized courses	7	88%	3.7%
Part of an introductory course	2	25%	N/A
No mandatory (specialized or introductory) course	0	0%	N/A
References in the other professional courses	N/a	N/a	N/A
Total (programs)	**8**		

Table 3. Reference to social / environmental responsibility in the educational programs in biomedical engineering (Russia)

There are only few, but large with respect to enrollment, relevant bachelor programs in Russia. Russian educational programs are regulated by the Federal State Educational Standard (FSES), with a few universities authorized to establish educational standards independently (only one for our list of programs). The earlier versions of FSES contained the list of competencies and educational outcomes along with the basic list of courses required in the curriculum. Consequently, all the examined programs contained similar “general cultural competencies”, such as “ability to develop one’s own world view basing on the philosophical knowledge” or “professional competencies”, such as “ability to monitor compliance with environmental safety”. In the latest version of FSES the list of competencies converged with the criteria of European Network for Accreditation of Engineering Education (EUR-ACE), which results in more concrete wording, e. g. “taking into account the economical, environmental and social constraints”.

In compliance with the previous version of FSES, all the programs included at least 3 Credits in *Philosophy* and 2 in *Ecology* (5 of 216 Credits). However, the institutions normally increased this share by adding more credits to both courses or by adding other courses, such as *Sociology*, *Philosophy of Technology* or *Environmental Monitoring*. This has led to relatively higher proportion of the responsibility-related courses (3.4%). Yet, the latest version of FSES contains no explicit requirement of the environmental course and one of the examined programs has already been updated to relocate this workload for “soft skills” training (foreign language and time-management) for the next year.

At least two programs address issues in ethics and risks in their introductory courses. Still, these introductory courses are typically very short (2 of 216 Credits).

Due to the lack of published syllabus, the non-specialized professional courses were not examined. 3 programs offer relevant electives in addition to the basic specialized courses, such as *Sociology* and *History of Science and Technology*.

Interpretation of the results and limitations

The analysis shows that the responsibility-related disciplines and subjects are underrepresented in the educational programs, in contrast to their declared objectives and outcomes. Although all the examined programs explicitly declare professional responsibility among their priorities, 17% of them contained no mandatory (introductory or specialized) course addressing social and environmental responsibility, safety or risks. As for the rest, the share of the workload devoted to these courses constituted only 2.5% on average, with little variation between the highly centralized and standardized educational system (Russia) and the system with high autonomy of the universities (the USA). According to the available syllabus, some institutions did formalize the responsibility agenda in the professional courses (up to 33% in the USA); however, the "ethics across the curriculum" approach has not been yet implemented consistently as a paradigm of a whole educational program. It is more likely to see occasional discussion of ethical and safety issues in a small number of professional courses within a program.

A few strategies of curriculum design are minimizing the workload devoted to this "impractical" part of engineering education. Firstly, the problems of professional impact of engineering can be addressed in the short introductory, sometimes even university-wide course (that is, before the students could face any particular professional challenge). Secondly, ethical, social and environmental assessment may be assigned as a task in the individual or group project courses – that is, with "ad hoc" approach, which seems to require students to "solve" it as a problem among others and without preliminary systematic instruction. Thirdly, the responsibility-related courses are sometimes positioned as mandatory but provide no credits, which may influence attendance. Finally, relevant courses are often found in the lists of electives. This seems to be the least desirable strategy due to the following considerations. The large institutions offer tens and hundreds of elective courses in social sciences and humanities, which may be extremely narrowly focused. It seems hard for responsibility-related topics to compete with more pragmatic courses, such as management, economics, psychology or "soft skills" development as well as with extremely attractive courses, such as "Philosophy of Food" or "Happiness in a Difficult World". More than that, it can be assumed that the very choice of a responsibility-related course implies at least preliminary awareness of the student, leading to reproduction of some kind of esoteric knowledge. In addition, the status of the electives (therefore, "optional" or "unnecessary" courses) may strongly influence the students' attitudes to the questions of professional responsibility.

Since only the actual available educational programs were considered, we have merely a simultaneous picture of the present-day educational policies. The sample for the USA was limited to 21 (out of more than hundred) accredited programs, which graduated more than 90 students in the last year (approximately 40% of the nation's total). Thus, the sample is non-representative, due to the possible peculiarities of organization in the largest universities and departments. Despite of the process of unification, credit count is still incompatible in different educational systems. Elaborating a universal unit of students' workload is a special methodological problem.

Conclusion

The study shows that, despite the differences in national and international standards, professional responsibility has become a declared outcome of biomedical engineering education at the undergraduate level in three different educational systems. The lively international methodological debate in the last two decades has suggested variety of ways to design curriculum to achieve this objective. Still, the content analysis of current educational programs revealed the underrepresentation of the responsibility-related disciplines and subjects. The existing strategies of economizing the most valuable time resource include "electivization" or substitution of the responsibility-related disciplines with more instrumental and popular courses. The curriculum study has its obvious limitations, giving merely a preliminary picture of the formally documented educational policies. The structure of curriculum reflects the state of institutionalization of educational practices. Still, the questions of content, quality, informal mechanisms and outcomes of teaching require further in-depth study and assessment by the stakeholders.

Acknowledgments

I thank Prof. Carl Mitcham (Colorado School of Mines) for stimulating ideas and Dr. Erik Fisher (Arizona State University) for constructive criticism. I thank the colleagues from Bauman Moscow State Technical University and the Institute for Technology Assessment and Systems Analysis of Karlsruhe Institute of Technology for support. I thank the organizers of Budapest Workshop on Philosophy of Technology 2019 for opportunity to present this study and the anonymous reviewers for their detailed comments.

References

Accreditation Board for Engineering and Technology. *Criteria for Accrediting Engineering Programs*. Baltimore: ABET, 2016. https://www.abet.org/wp-content/uploads/2016/12/E001-17-18-EAC-Criteria-10-29-16-1.pdf

Accreditation Board for Engineering and Technology. *Criteria for Accrediting Engineering Programs*. Baltimore: ABET, 2017. https://www.abet.org/wp-content/uploads/2018/02/E001-18-19-EAC-Criteria-11-29-17.pdf

Beder, Sharon. *The New Engineer: Management and Professional Responsibility in a Changing World*. South Yarra: Macmillan Education Australia, 1998.

Bucciarelli, Louis L. "Ethics and Engineering Education." *European Journal of Engineering Education* 33 no. 2 (2008): 141–49. https://doi.org/10.1080/03043790801979856.

Bucciarelli, Louis L., and David E. Drew. "Liberal Studies in Engineering – a Design Plan." *Engineering Studies* 7 no. 2-3 (2015): 103–22. https://doi.org/10.1080/19378629.2015.1077253.

Christensen, Steen Hyldgaard, and Erik Ernø-Kjølhede. "Epistemology, Ontology and Ethics: 'Galaxies Away from the Engineering World'?" *European Journal of Engineering Education* 33 no. 5-6 (2008): 561–71. https://doi.org/10.1080/03043790802568070.

Christensen, Steen Hyldgaard, Martin Meganck, and Bernard Delahousse. "Introduction. Occupational building in engineering education." In *Philosophy in Engineering*, edited by Steen Hyldgaard Christensen, 13–22. Aarus: Academia, 2007.

Conlon, Eddie. "The New Engineer: between Employability and Social Responsibility." *European Journal of Engineering Education* 33 no. 2 (2008): 151–59. https://doi.org/10.1080/03043790801996371.

Date, Geetanjali, and Sanjay Chandrasekharan. "Beyond Efficiency: Engineering for Sustainability Requires Solving for Pattern." *Engineering Studies* 10 no. 1 (2017): 12–37. https://doi.org/10.1080/19378629.2017.1410160.

Downey, Gary Lee. "What Is Engineering Studies for? Dominant Practices and Scalable Scholarship." *Engineering Studies* 1 no. 1 (2009): 55–76. https://doi.org/10.1080/19378620902786499.

Downey, Gary Lee. "Opening up Engineering Formation." *Engineering Studies* 7 no. 2-3 (2015): 217–20. https://doi.org/10.1080/19378629.2015.1121612.

Mitcham, Carl, and Elaine E. Englehardt. "Ethics Across the Curriculum: Prospects for Broader (and Deeper) Teaching and Learning in Research and Engineering Ethics." *Science and Engineering Ethics* 25 no. 6 (December 2019): 1735–62. https://doi.org/10.1007/s11948-016-9797-7.

Mitcham, Carl. "A Historico-Ethical Perspective on Engineering Education: from Use and Convenience to Policy Engagement." *Engineering Studies* 1 no. 1 (2009): 35–53. https://doi.org/10.1080/19378620902725166.

National Academy of Engineering. *Educating the Engineer of 2020: Adapting Engineering Education to the New Century*. Washington, DC: The National Academies Press, 2005. https://doi.org/10.17226/11338.

Pinch, Trevor. "Teaching sociology to science and engineering students: some experiences from an introductory science and technology studies course", In *Integrating the Sciences and Society: Challenges, Practices, and Potentials*, edited by Harriet Hartman, 99-114. Bingley: Emerald Group Publishing Limited, 2008. https://doi.org/10.1016/S0196-1152(08)16005-7.

Prasad, Jitendra, Avijit Goswami, Brijesh Kumbhani, Chittaranjan Mishra, Himanshu Tyagi, Jung Hyun Jun, Kamal Kumar Choudhary, et al. "Engineering Curriculum Development Based on Education Theories." *Current Science* 114 no. 09 (2018): 1829-34. https://doi.org/10.18520/cs/v114/i09/1829-1834.

Stephan, Karl D. "A Survey of Ethics-Related Instruction in U.S. Engineering Programs." *Journal of Engineering Education* 88 no. 4 (October 1999): 459–64. https://doi.org/10.1002/j.2168-9830.1999.tb00474.x.

The Royal Academy of Engineering. *Educating Engineers for the 21st Century.* London: The Royal Academy of Engineering, 2007. https://www.raeng.org.uk/publications/reports/educating-engineers-21st-century

The United Nations Scientific, Educational and Cultural Organization. *Engineering: issues, challenges and opportunities for development; UNESCO report.* Paris: UNESCO, 2010. https://unesdoc.unesco.org/ark:/48223/pf0000189753_eng

Author information

Aleksandra Kazakova, Gubkin Russian State University of Oil and Gas (National Research University), Bauman Moscow State Technical University (National Research University) https://orcid.org/0000-0002-2952-8373

When AIs Say Yes and I Say No: On the Tension between AI's Decision and Human's Decision from the Epistemological Perspectives

Chang-Yun Ku

Abstract

Let's start with a thought experiment. A patient is waiting in the clinic room for the diagnosis result to decide whether he needs brain surgery for his medical conditions. After SaMD processed, the result shows that the patient is classified into the high-risk group with 99.9% of death rates and needs brain surgery immediately. But the result is opposite to your diagnosis that the patient needs not the surgery. Will you, as a physician in this scenario, object the result that SaMD has made?

Theoretically, Human should be the one who determines all the decisions and takes AI's results for reference only, as the GDPR Article 22 presumes. But quite the opposite, AI's result has greater influences on Human than we thought. In this paper, I explore the tension between AI's decision and human decision from the Epistemological perspectives, i.e. to justify the reasons behind the positive human beliefs in AI. My conclusion is that positive human beliefs in AI are because we misidentified AI as a general technology, and only if we can recognize their differences correctly, then the requirement of "Human in the loop" in the GDPR Article 22 can have its meaning and function.

Keywords: Artificial Intelligence, GDPR Article 22, Human in the Loop, Automated Decision-making

Introduction

U.S. FDA (U.S. Food & Drug Administration) approves SaMD (Software as Medical Device)[1] for medical diagnosis. China uses the AI social credit system Zhima Credit[2] to replace the traditional financial credit score. Estonia is going to deploy Robot Judge in Court for the small claim cases (Niiler 2019).[3] These examples from different Countries in different fields show that the Artificial Intelligence (AI/AIs), or say the algorithm, is not only an idea in the Sci-Fi but also a reality in our daily lives. Not even mention those AI applications in the private sectors. And Article 22 of the GDPR (EU General Data Protection

[1] U.S. Food & Drug Administration. "Software as a Medical Device (SaMD)." Accessed February 8, 2020. https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd

[2] ZHIMA Credit, Ant Financial Services Group. Accessed February 8, 2020. https://www.xin.xin/#/home

[3] Niiler, Eric. "Can AI Be a Fair Judge in Court? Estonia Thinks So." WIRED. March 25, 2019. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/

Regulation)[4] on "*Automated individual decision-making, including profiling*" already regulated that for those decisions that are seriously impactful to the data subject should not be determined solely by automated decision-making process.

But the thing isn't as perfect as it sounds. Three empirical studies show that AI's influences on Human are more than we thought. Results from AI decision-making are better than we humans do? If not, what's the reason that we humans believe in AI's decision than Human? To explore this critical issue, I divide this paper into 7 parts. First, I will start from the elaboration of the GDPR Art.22 to point out the importance of "human in the loop" in the automated decision-making process. Second, I will take three experiments' results to show the significant impact of AI to Human, when human facing decision-making process. Third, I'll prove that the presumption of "AI's decision is better than human decision" isn't solid by demonstrating the nature of AI's decision. After these, I will point out the misattribution of AI as a general Technology is the reason we human have believed in AI. Fifth, I'll return to the Art.22 of GDPR and propose three possible solutions from epistemological perspectives to resolve the gap between this provision and the reality. Sixth, I'll push the discussion further for the crucial issue of whether the expert is immune to AI's decision when making professional decisions. And finally, I will summarize my arguments and conclude this article.

The Requirement of "Human in the Loop" in the GDPR Article 22

The GDPR recognized that "… profiling and automated decision-making can pose significant risks for individuals' rights and freedoms which require appropriate safeguards" (WP29 2018)[5], and thus it regulated automated decision-making process in the Art.22 titled *Automated individual decision-making, including profiling.* The contents of this provision are as followed:

1. *The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*
2. *Paragraph 1 should not apply if the decision:*
   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   (b) is authorized by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subjects rights and freedoms and legitimate interests; or
   (c) is based on the data subject's explicit consent.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[5] Article 29 Data Protection Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01). 2018. P.5.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decision referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

According to the Art.22, basically, the "solely automated processing" that could cause "legal effects" or "similarly significant effects" to the data subject, is prohibited. The "Solely" automated processing means the processing is "without human intervention", i.e. the result of automated processing was decided by the algorithm and then automatically delivered to the data subject, but with no prior or meaningful assessment by a human (WP29 2018).[6]

The "legal effects" means that a decision affects someone's legal rights or legal status[7], and the WP29 (Article 29 Data Protection Working Party) names a few examples as the legal right and the legal status. The Legal rights include the freedom to associate, to vote or to take a legal action; and the legal status includes cancellation of a contract, denial of social benefit, refused admission to a country…etc. And the term "similarly significantly effects", refer to the results that are serious impactful to the data subject and thus require the protections under this provision (WP29 2018)[8]; although it's not directly defined in the GDPR, the WP29 explains this as an effect that "must be similar to that of a decision producing a legal effect" (WP29 2018)[9], for example affect someone's financial circumstances, access to health service or employment opportunity…etc. In other words, if the effect of solely automated decision-making isn't serious impactful, then it will not to be regulated or prohibited by the GDPR Art. 22.

Even if the process could cause the legal effect or the similarly significant effect, but under three specific conditions, i.e., for the contract, with the Union or Member State's authorization, or with data subject's explicit consent, the GDPR allows the use of solely automated individual decision-making process. The permissions are under the conditions that if the data controller can meet the requirement of providing appropriate safeguards. These appropriate safeguards include data controller needs to provide meaningful information, specifically the logic of automated decision-making process, to the data subject (WP29 2018)[10], and also provide the opportunities for the data subject to request human intervention, to contest the decision, and to obtain the explanation (GDPR Recital 71)[11].

As mentioned above, for the automated decisions-making process that is regulated by the GDPR Art.22, first, it's only limited to those decisions will cause the significant effect to the data subject in principle; Second, with suitable measures, i.e. human interven-

[6] WP251, p.9.

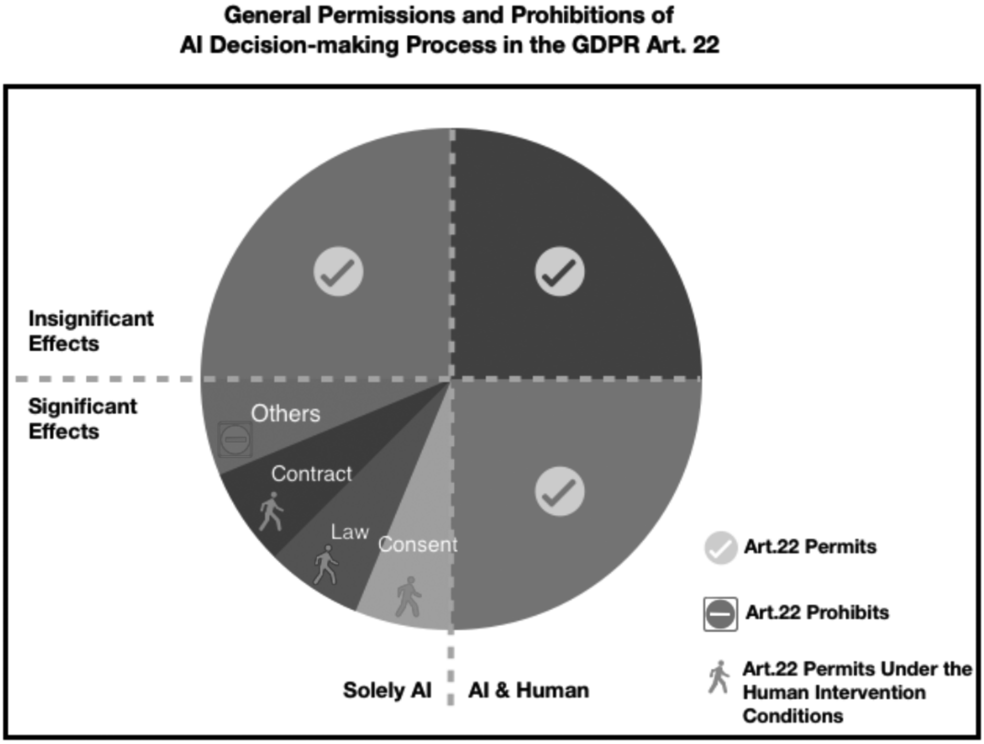
[7] WP251, p.21.

[8] WP251, p.22.

[9] WP251, p.21.

[10] WP251, p.20.

[11] GDPR Recital 71


General Permissions and Prohibitions of
AI Decision-making Process in the GDPR Art. 22
Insignificant Effects
Significant Effects
Others
Contract
Law
Consent
Solely AI
AI & Human
Art.22 Permits
Art.22 Prohibits
Art.22 Permits Under the Human Intervention Conditions

tions, the GDPR permits to use solely automated decision-making for three specific purpose. The "human factor" here is a means to prevent the harm that a solely automated individual decision-making can cause. And this Human is expected to oversee the decision, and who must "… has the authority and competence to change the decision" and "consider all the relevant data" (WP251 2018)[12] In the GDPR's presumption, the Human is capable of making things right when AI goes wrong, and has the authority to determine the best possible decision with AI's advice.

The Inconvenient Truth: The Tension between AI's Decision and Human Decision

To avoid the possible harms, the GDPR Art.22 regulates solely automated decision-making process, which could cause significant effects to the data subject, by requiring human intervention. "Human in the loop" is the solution to the risks post by solely automated decision-making process, i.e. makes the solely automated process "not" solely. But, does this solution actually work? In this section, I would like to introduce three empirical studies to point out a surprising phenomenon: AI 's decision has more influence on Human than we could image, and the Human is actually leaded by AI.

[12] WP251, p.21.

First of all, Human seems more than willing to take the "advice" from AI. Logg et al.'s (Logg et al. 2019)[13] research started from the prevalent presumption of "algorithm aversion", a term refers by Dietvorst et al. (Dietvorst et al. 2015)[14] means that humans distrust algorithm even though algorithm consistently outperform humans. Logg et al. 's study designed to enquire "the role of the self", when human facing the advice from the algorithm, from other people or from people themselves. According to their research results, when people with a choice to take advice from themselves or from other people, 88% of participants would take their own advice. But when people can choose the advices from themselves or from the algorithm, 66% people would take algorithm's advice instead of their owns, even Logg et al. specifically claimed that "the model does not have any additional information that you will not receive" to the research participants (Logg et al. 2019)[15]. The results of their experiments show that "people readily rely on algorithmic advice" (Logg et al. 2019)[16], and Logg et al. call this phenomenon "algorithm appreciation".

Second, the algorithmic advice has a strong impact to human decision than we knew. Vaccaro and Waldo's used the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk scores to test the effects of algorithmic results to human assessments (Vaccaro and Waldo 2019).[17] They divided research participants into two groups: One group was given 40 defendants' profiles with low-risk score, and the other group received the same 40 defendants' profiles but with high-risk score. Their study results show that the scores in high-risk scores group is in average 42.3% higher than the scores in low-risk score group, i.e. AI's results effect people's decision significantly. And this result confirmed the hypothesis of psychological cognitive bias in the Human—the "anchoring effect" —when using algorithmic predictions. Anchoring effect means that during human assessments process, the algorithmic result will "act as an anchor" and thus "individuals will assimilate their estimates to a previously considered standard" (Vaccaro and Waldo 2019)[18]. Vaccaro and Waldo concluded that "even if algorithms do not officially make decisions, they anchor human decision in serious way" (Vaccaro and Waldo 2019).[19]

Thirdly, the Human generally believes algorithmic prediction, even if the accuracy is no more than we flip a coin. Lai and Tan conducted the experiments to evaluate humans' performance when people use different levels of machine assistance, and to see the influ-

[13] Logg, J. M., J. A. Minson and D. A. Moore. "Algorithm appreciation: people prefer algorithmic to human judgment." Organizational Behavior and Human Decision Processes, Vol.: 151 (February 5, 2019): 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

[14] Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." Journal of Experimental Psychology: General, Vol. 144, Issue 1 (February 2015): 114-126. https://doi.org/10.1037/xge0000033
P.114.

[15] Logg et al. 2019, p.96.

[16] Logg et al. 2019, p.99.

[17] Vaccaro, Michelle and Jim Waldo. "The Effects of Mixing Machine Learning and Human Judgment." Communications of the ACM Vol. 62, No.11 (October, 2019): 104 -110. https://doi.org/10.1145/3359338.

[18] Vaccaro et al. 2019, p.108.

[19] Vaccaro et al. 2019, p.105.

ences of machine accuracies to human predictions (Lai and Tan 2019).[20] They designed different levels of machine assistance for research participants, and they found that machine results with the description of "predicted label", i.e. attached the description of " the machine predicts…" to the results, could effectively improve research participants' performances than without it. Lai and Tan pointed out "this observation also echoes with concerns about humans overly relying on machines" (Lai and Tan 2019).[21] Also, Lai et al.'s research results showed that the participants' trusts are effected by the machine accuracies; more precisely, when the machine accuracies downed from 87%, 70%, 60% to 50% of machine predictions, the degrees of human trusts decreased from 79.6%, 78.6%, 76.9% and 74.5% accordingly. With these insignificantly incline results between the degrees of machine accuracies and the degree of humans' trusts, Lai and Tan concluded that "our findings suggest that any indication of machine accuracy, be it high or low, improves human trust in the machine" (Lai and Tan 2019).[22]

These studies above disclose the inconvenient truth that Human is greatly influenced by the so-called AI, the machine, or the algorithm. AI's result has the essential power to impact human judgment psychologically, whether it works as an anchor and makes humans adjust their decision toward it, or it could even make humans give up their chances to decide completely by taking AI's result instead. Even humans know the accuracies of AI are no more than we toss the coin, i.e. the chances of correctness are 50/50, humans still rather take AI's result anyway.

Following from this inconvenient truth, when the differences or conflicts emerge between the human decision and AI's decision, Human has high possibilities to assimilate human decision toward to AI's decision, or take AI's decision as the final decision instead. In other words, when Human prefers to follow AI's decision, it will eventually lead human to yield the decision authority to AI, as long as AI is in the decision loop. And this phenomenon also makes the GDPR Art. 22 meaningless, i.e. even with the requirement of "human in the loop" to prevent the harms that solely automated process could cause to Human, the human decision-maker will neither choose a different direction from AI, nor challenge AI's result.

And according to this human preference, it's reasonable to ask, why human generally inclines to believe AI's result but not human result? For this question, we are actually asking why Human has these great positive beliefs in AI, or why we believe the result or advice that AI bringing to us have more value or even closer to the truth? And from the epistemological perspectives, more importantly, is the reasons behind this "human positive beliefs" justified? To answer these questions, we need to inquire into two topics sequentially, i.e. the nature and the difference of AI's decision, and the formation of human positive beliefs in AI.

[20] Lai, Vivian & Chenhao Tan. "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection." In FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 29-38. United States, New York: Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287590.
[21] Lai and Tan 2019, p.34.
[22] Lai and Tan, p. 36.

The Nature and the Difference of AI's Decision

Why does Human believe AI's decision or AI's result has more value or closer to the truth than human decision or human result? The presumption of this "positive beliefs" is reasonable if and only if humans presume "AI's decision is definitely better than human decision". Is this presumption true? The only way to justify it is to inquire about the nature of AI's decision and the difference between AI and Human decisions.

The Artificial intelligence is based on the models of Machine Learning (ML) and its branch Deep Learning (DL), and the algorithm is the core to perform these functions. And the material for ML is data, or says Big Data, whether or not it's personal data. The Volume, Variety, Velocity of data are much more than a human can perceive, and thus these data is considered by Human has the qualities of Veracity, Variability and Value. But when the issue comes to AI decision-making, the training data for AI is based on those past decisions that humans have made.

Regards to AI's performance of decision-making, if the training data is generated from those decisions that humans used to make, then, I believe that AI's decision will basically repeat the human decision, or could get even worse, for the five characters that can derive from the training data. These five characters are as follows.

The first character is that AI will repeat the human choice, no matter that decision is right or wrong. What we used to choose, AI will choose the same; what we have decided in the past, AI will have the same decision in the future. And based on this correspondence between the training data and the AI's performance, AI will surely make the same wrong decision as what Human did before. Furthermore, as G. Marcus points out, "Human beings can learn abstract relationship in a few trials…on a capacity to represent abstract relationships between algebra-like variables…; Deep learning currently lacks a mechanism for learning abstractions through explicit, verbal definition, and works best when there are thousands, millions or even billions of training examples…(Marcus 2018)"[23], when facing an unexperienced event, i.e. the new condition is nothing in common with the data in the training datasets, the performance of AI will be no better than Human do.

The second is that AI will amplify the injustice of human past decisions, even though it was unnoticed before. Because of the volume of the training data, AI will not only repeat human past mistakes, but it'll amplify the human errors as well. And the bias we wrongly formed for a decision in the past will become AI's dominant criterion for a decision now. For example, Amazon's AI recruiting tool is abandoned because this AI tool prefers the male candidate than the female candidate, and the reason is the training data of last ten years shows the company hired more men than women (Dastin 2018)[24]; and V. Ordóñez and his team discovered the gender bias is performed when depiction of activities by ML research–image collections, e.g. the activity of washing is linked to women, and the coaching activity is linked to men, since the training data generally linked these activities to the certain gender (Simonite 2017)[25]. As M. Yatskar said "this could work to not only reinforce

[23] Marcus, Gary. "Deep Learning: A Critical Appraisal." ArXiv abs/1801.00631 (January 2018).

[24] Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women." Reuters, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

existing social biases but actually make them worse"(Simonite 2017)[26]. The hidden unjust factor of a human decision is now the main criteria of AI's decision.

The third character is that AI will miss the truly but "off the record" factors. Even if we consider every visible factor of a human decision, but sometimes, the true cause that a human decision-making based on is the "off the record" one, i.e. the factor literally isn't or couldn't be recorded into the training data. The human decision could base on empathy, intuition, memories or any other different kinds of "human things", and these factors aren't possible to be included into the records or the training dataset. "Human judgment is affected by a range of invisible factors that the decision-maker is unable to fully explain when scrutinized" as A. Babuta descripts, the influence of the "Noise" to human decision-making are usual overlooked (Babuta 2018).[27]From this point of view, the training data for AI isn't accurate, and so does the decision that AI makes.

The fourth one is that AI will calculate more factors than a decision need. The algorithm could wrongly link the similar but irrelevant factors of each decision, and thus AI will be trained to make a decision based on those irrelevant factors, e.g. the correlation problem. In the famous Gettier's Problem[28], although the "ten coins" factor has nothing to do with the decision that which job candidate will be hired, but once the former is linked to the latter by algorithm, as what Smith did, AI will definitely decide the later by the former factor, i.e. hire the person only because who has ten coins in the pocket. And as W.T. Chiou indicates "Correlation knowledge is invaluable as it bridges the information gaps and allow a decision to be made in case of ignorance" (Chiou 2018)[29], but "Without causal explanations, decisions based on mere correlations would amount to arbitrary actions that would blame an affected subject for something that cannot be attributable to him or her"(Chiou 2018)[30].

Fifth, it's acknowledged that there are few kinds of AI's decision results are totally disasters, e.g. the AI's face recognition function. The highly inaccuracy of Facial-recognition AI's has widely reported in the media, and the latest NIST's (National Institute of Standards and Technology, U.S.) research result confirmed that "we found empirical evidence for the existence of demographic differentials in the majority of contemporary face recognition algorithms that we evaluated" (NIST 2019)[31]. According to NIST's report, Asian and American Indian individuals have a higher false negative rate than races, and Woman has 2 to 5

[25] Simonite, Tom. "Machines Taught by Photos Learn a Sexist View of Women." WIRED, August 21, 2017. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

[26] Simonite 2017.

[27] Babuta, Alexander. "Innocent Until Predicted Guilty? Artificial Intelligence and Police Decision-Making." RUSI Newsbrief Vol. 38, No. 2(March 2018). https://rusi.org/sites/default/files/20180329_rusi_newsbrief_vol.38_no.2_babuta_web.pdf

[28] Gettier, Edmund L. "Is Justified True Belief Knowledge?" Analysis, Vol. 23, Issue 6 (June 1963): 121–123. https://doi.org/10.1093/analys/23.6.121

[29] Chiou, Wen-Tsong. "Causal Explanation as a Partial Solution to Algorithmic Harms." paper presented at The 7th Academia Sinica Conference on Law, Science and Technology: Emerging Legal Issues for Artificial Intelligence: Legal Liability, Discrimination, Intellectual Property Rights and Beyond, Taipei, Taiwan, November 26-27, 2018, 1-16.

[30] Chiou 2018, p.8-14.

[31] National Institute of Standards and Technology, U.S. Department of Commerce. "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects." (December 19, 2019) P.7. https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

times of false positives rates higher than man (NIST 2019)[32]. The reason to cause the inaccuracy of Facial-recognition AI could be many, but as IBM points out "One of the biggest issues causing bias in the area of facial analysis is the lack of diverse data to train systems on" (IBM 2018)[33]. The San Francisco City is the first city in the U.S. passed the Acquisition of Surveillance Technology Ordinance to ban the local agencies for using AI's facial recognition technology in May 2019 (Board of Supervisors, San Francisco 2019)[34], and many other States in the U.S. like New Hampshire, Washington, Indiana, South Carolina etc. are proposing different level's restrictions to Facial-recognition AI (Ramos & Abernethy 2019)[35].

As demonstrated above, in comparing the AI's decision to Human decision, there is no guarantee that AI's decision will definitely better than human decision. In fact, from the training data perspective, AI could precisely repeat human decision and amplify the unjust one, but the flexibility of Human to adjust the decision from the aware mistakes in the past is absent in AI. Human ability of adjustment is the answer to why AI's decision is no better than Human.

Of course, for the record, I have no intentions to deny the possibility that AI could bring a better decision than Human do. But from the Philosophical perspectives, the reasoning is as important as the conclusion, and sometimes it's even more important than the result. Thus, these factors that AI computes for a specific decision should be elaborated and disclosed, as The Right to Explanation or the Algorithmic Transparency that many scholars request. Otherwise, how can we prove that the AI's decision is truly better than Human? And these requirements aren't seem to be achievable by AI in the present days.

Our Beliefs in AI: An Empirical Explanation

People are holding positive and optimistic attitudes toward AI. According to ARM Northstar's survey which across the U.S., E.U. and Asia, 61% of 3938 participants believe AI will make the world a better place, especially in healthcare, science and traffic control fields (ARM Northstar 2017)[36]. As Genesys' National online survey, it shows that 70% of 1001 U.S. employees are with the positive attitude toward AI's impact of workplace (GENESYS 2019)[37]. And similarly, according to the Northeastern University and Gallup's mail survey

[32] NIST 2019, p.7-8.

[33] IBM. "IBM to release world's largest annotation dataset for studying bias in facial analysis." Accessed February 8, 2020. https://www.ibm.com/blogs/research/2018/06/ai-facial-analytics/

[34] Board of Supervisors, City and County of San Francisco. Administrative Code-Acquisition of Surveillance Ordinance. Accessed February 9 2020. https://sfgov.legistar.com/LegislationDetail.aspx?ID=3953862&GUID=926469C0-A7BA-47D3-BB32-05C2C6D8EB2B

[35] Ramos, Gretchen A. and Darren Abernethy. "Additional U.S. State Advance the State Privacy Legislation Trend in 2020." The National Law Review. January 27,2020. https://www.natlawreview.com /article/additional-us-states-advance-state-privacy-legislation-trend-2020

[36] ARM and Northstar. "AI today, AI tomorrow: Awareness, acceptance, and anticipation of AI: A global consumer perspective." 2017. http://pages.arm.com/rs/312-SAX-488/images/arm-ai-survey-report.pdf

[37] GENESYS. "70% of U.S. Employees Hold Positive View of Artificial Intelligence in the Workplace Today. " July 10, 2019. Accessed February 8 2020.
https://www.prnewswire.com/news-releases/70-of-us-employees-hold-positive-view-of-artificial-intelligence-in-the-workplace-today-300882125.html

of 3297 adults in the U.S., 76% participants agree or strongly agree that AI will change the ways people work and live in the next 10 years, and 77% of them are mostly positive or very positive about the impact that AI will bring (Northeastern University and Gallup 2018).[38]

But, as mentioned previously, the presumption of "AI's decision is definitely better than human decision" is not true, and AI's decision is basically equal or worse to Human. So, it's naturally to ask why Human would choose AI's decision instead of Human decision, why Human has more beliefs in AI than Human, and where are these positive beliefs in AI coming from? In short, how did Human form these positive beliefs in AI?

We always have positive impressions attached to the term "Technology", and also attribute it with great characters, like efficiency, accuracy, convenience, cost saving etc. Our beliefs of these characters belong to the "Technology" are gradually forming from our experiences, i.e. our past interactions with these Technological products. Thus, we always have similar expectations of those Technological Inventions, as long as they are in relation to Technology. And AI is presumed by Human to be one of those general Technological Inventions.

But, AI and the "general Technology" are heterogeneous. As Artificial Intelligence is deemed to play the leading role in "The Fourth Industrial Revolution", the term was introduced by the Founder and the Executive Chairman of World Economic Forum K. Schwab (Schwab 2015)[39] this should be the signal that AI is naturally different from the ordinary technology that we used to know. When refer to the decision-making, from the epistemological perspectives, the genuine difference between AI and general Technology is the certainty of the judging criteria that are used for decision-making process.

AI is expected to learn and generate the mean standards from training data instantly and apply them promptly, but the general Technology is designed to follow a specific standard and apply it afterwards. The new data input will definitely influence the mean standard of the training dataset, and consequently affect the judging criteria for AI decision-making. The fluctuation of mean standards causes AI's result not only unpredictable, but also makes AI's decision lack of stability and even possible inconsistency. For this reason, the epistemic certainty of the AI's judging criteria and decision results differs from the general Technology. Thus, it's by no means that we could categorize AI as a "general technology" from this viewpoint. But nonetheless, we still give AI all the credit and beliefs of the Technology as usual, like AI is a kind of general or ordinary technology.

A belief needs to be justified to become a truth that is worth to believe, and knowing is the justification of a belief. Surely, a true belief without justification can be acquired accidentally, i.e. by luck, but that's not the way we Humans would expect, because the "luck" is unreliable. Thus, to know before to believe, is the fundamental principle for us to see the World outside of us. Even if "to know" can have a different meaning or different degree, but the "degree of knowing" should be as equal as possible to the "degree of belief".

[38] Northeastern University and Gallup Inc.. "Optimism and Anxiety: Views on the Impact of Artificial Intelligence and Higher Education's Response." October22, 2018. https://perma.cc/57NW-XCQN

[39] Schwab, Klaus. "The Fourth Industrial Revolution: What It Means and How to Respond. Foreign Affairs." Foreign Affairs. December 12 2015. Accessed February 8, 2020. https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution

But it seems like we humans don't know what AI is yet. Pegasystem conducted a survey of 6000 adults in North America, APAC and EMEA, 70% of participants believe they understand AI, but 50% of them don't understand AI can solve problems, 37% of them don't understand AI can interpret speech, and 35% of them don't understand AI can mimic humans … etc. (Pegasystem 2017).[40] And according to Entrata's online survey result from 1051 U.S. participants, over 38% just heard of AI or have no idea what it is, but 52% of participants are comfortable interacting with it (Entrata 2019).[41] As Bristows's surveyed in U.K. shows that in 2103 participants, 25.5% of all never heard of AI or have heard the term but unsure what it is, and 39.5% participants says has limited knowledge of it (Bristow 2018).[42] The underlying problem of this is, as I. Evans highlighted that "Nobody agrees on what AI is" (Evans 2019)[43] and explained in Elsevier' report "The AI field has multiple definitions, but lacks a universally agreed understanding. AI means different things to different people: there are more differences than commonalities… ", (Elsevier 2019)[44] there is no one definition of AI that can achieve human consensus, i.e. none of these definitions can explain AI' characters, abilities, functions, and potentials in the complete, clear and distinct ways.

With no definitions or explanations of AI can satisfy with these basic criterion of understanding, but only based on what we have known by now, Humans have given more beliefs to AI than it should have. Our beliefs in AI shouldn't be the same as the general technology that we used to know, even if AI is included in the broadest definition of Technology. AI should be considered separately from the general or ordinary Technology.

Rethinking GDPR Article 22: The Possible Solutions

In the GDPR Art. 22, it requires human intervention as the safeguards to prevent the harms that the solely automated decision-making could cause. But humans prefer to choose AI's decisions or results instead of Humans decisions, due to we simplified the nature of AI as a kind of general technology. Following from this epistemological conclusion, I would like to propose three possible solutions to resolve this conflict between the regulation and the reality.

[40] Pegasystem."What Consumers Really Think About AI: A global Study." June 19, 2017. https://www.pega.com/insights/resources/what-consumers-really-think-ai-infographic

[41] Entrata. "Artificial Intelligence and Apartment Living: Survey Studies Consumer's Knowledge of and Attitude Toward AI (Report)" and " What Consumers Really Think About AI (poster)". August 2019. P.18. http://info.entrata.com/newsletters/case_studies/AI/SurveySummary.pdf

[42] Bristows. Artificial Intelligence: Public Perception, Attitude and Trust. 2018. P.7. https://d1pvkxkakgv4jo.cloudfront.net/app/uploads/2019/06/11090555/Artificial-Intelligence-Public-Perception-Attitude-and-Trust.pdf

[43] Evans, Ian. ""Nobody agrees on what AI is"- How Elsevier's report used AI to define the undefinable." Elsevier. January 18, 2019. Accessed February 8, 2019. https://www.elsevier.com/connect/nobody-agrees-on-what-ai-is-how-elseviers-report-used-ai-to-define-the-undefinable

[44] Elsevier. "Artificial Intelligence: How knowledge is created, transferred, and used." January 2019. https://www.elsevier.com/research-intelligence/resource-library/ai-report?utm_source=AI-EC

The first is to raise public awareness of AI. Since no sufficient knowledge of AI and misunderstand AI's nature are the reasons why humans prefer to follow or obey AI's decision, as I explained in the previous section, then, to clarify AI's characters to the general public should be the first thing to do. The Government has more resources and more power than any other private sector, thus should be the one who carries this responsibility to raise public awareness of AI.

The second is the mechanism of the third party certification for AI's neutrality. The rationale behind this mechanism is as follows: if an organization or a company designs AI, they may not be able to verify the result objectively; and if they can't verify the results objectively, then the non-neutral results might affect the decisions of their employees, who prefer to choose AI's decision as described previously; Thus it's important to keep AI as neutral as possible, and maybe the mechanism of the third party inspection and certification could help. This mechanism is like those existing third party inspection methods, but its purpose is for examining AI's neutrality. This third party inspection agency could be a government agency or a private company, as long as it has the Governments' license. This mechanism should be applied before the AI's actual application and periodic inspection afterward.

Last but not least, it's necessary for us to clarify what do Humans really want from AI, and the reasons for AI's developments should be more than efficiency and resource saving. Everyone expects AI can improve our lives dramatically, and everybody talks about the benefits that AI could bring to the Humans. In the meantime, Humans shouldn't forget how little we know about the AI, and those unforeseen or unknown consequences due to the limited knowledge of AI we have. While we devote ourselves to improving AI's applications, we should enquire the meaning of this dedication as well.

Further Discussion: The Dichotomy of the Layperson and the Expert

Maybe, the optimistic attitudes toward AI or the willingness to embrace AI's decision don't happen to everyone? One important topic also studied by Logg et al. is the different attitudes between the layperson and the expert, when they're provided with algorithmic advice at the decision-making moment. Logg et al. invited experts to predict the events in relation to their expertise, and the research results showed that experts "… adherence to their prior judgments and their failure to utilize the information offered to them ultimately lowered their accuracy relative to the lay sample"[45], Logg et al. further pointed out that this could be used to explain P.E. Meehl's theory that "why pilots, doctors, and other experts are resistant to algorithmic advice."[46]. According to this research result, the experienced experts seem to refuse AI's decision while making decisions in relation to their professions, and thus don't even consider AI's decisions. In short, the experts seem immune to AI's influence in their professions.

Whether there is a dichotomy of the layperson and the expert concerning the influence of AI's decision is crucial in two ways. First, this claim seems intuitive, for instance,

[45] Logg et al. 2019, p.99.
[46] Logg et al. 2019, p.99.

an experienced driver knows the routes and the traffic situations at a certain timing in general, therefore the driver won't need the GPS's advice. Second, if there is clear evidence can support the claim that the experts are immune to AI's decisions, then, there is no need to worry for those experts will be lead by AI or won't challenge to AI's decision in their professions, while they are making those professional decisions that are seriously impactful to the data subjects.

But interestingly, we can see the research results that include both inclinations, i.e. does and doesn't support the claim that "expert immune from AI's decision"; and when given similar conditions and the same professions, this contradictory is even clear. For example, M. Stevenson[47] analyzed the criminal court's data in Kentucky U.S. to see the outcomes after the State has mandated the use of the pretrial risk assessment algorithm. According to Stevenson's results, the judges did use the risk assessment as the State required. And from the release rate perspective, although it increases low-risk and moderate-risk rate in 22% and 16% of non-financial release, it also caused the non-financial bond change to released on the low cash bond, and Stevenson pointed out "thus, the net effects on the release rate were attenuated"(Stevenson 2018)[48]. As for the changes in total release rate, this mandatory increased low and moderate risk defendants rate in 9% and 7%, but also decreased 4% in high-risk defendants release rate, and Stevenson concluded that "in total, this resulted in a 4 percentage point increase in the release rate for all defendants, which eroded over time as judges returned to their previous bail setting habits"(Stevenson 2018)[49]. According to Stevenson's research result, even provided with the algorithmic advice, judges eventually use their experience as the guidance for decision-making.

On the other hand, B. Cowgill's research suggested a different perspective in the same professions (Cowgill 2018).[50] To research the issue of judicial compliance of algorithmic risk assessments, Cowgill analyzed the data from the criminal court of Broward County Florida U.S. As Cowgill's research results, the algorithmic advice increase the pre-trial detention for one week in average; and for the overall day in jail, the low/medium risk of general recidivism increase two weeks additional detention, and it's even double up for the violent recidivism (Cowgill 2018)[51]. According to the research results, Cowgill pointed out "the algorithmic guidance does affect pretrial bail decisions"(Cowgill 2018)[52], and further indicated that, in summary, "this result suggested that algorithmic suggestion have a causal impact on criminal proceedings and recidivism"(Cowgill 2018)[53]. As Cowgill's research result, algorithmic advice is the guidance for judges' decision-making process.

By these two research results that have contrary conclusions but are in the same profession, the experts in their professions will resist and ignore the AI's decision, or will follow and obey AI's decision, from the Epistemological perspective, I believe we should

[47] Stevenson, Megan T.. "Assessing Risk Assessment in Action." 103 Minnesota Law Review (2018): 303-384. http://dx.doi.org/10.2139/ssrn.3016088

[48] Stevenson 2018, p. 368.

[49] Stevenson 2018, p. 369.

[50] Cowgill, Bo. "The Impact of Algorithms on Judical Discretion: Evidence from Regression Discontinuities". (Working Paper)(December 5, 2018).
http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf

[51] Cowgill 2018, p. 11-12.

[52] Cowgill 2018, p. 12.

[53] Cowgill 2018, p. 1.

suspend our judgments on this issue temporarily, because we need more information and further researches, before we firmly claim that the experts refuse or decline of AI's decision in their professional decision-making process.

But, suspend our judgments on the "experts immune to AI's decision" issue doesn't mean we should just wait until the decisive results, and then to see if we need to take any action afterward. As the ultimate purpose of the Epistemology is to take the action in accordance with our knowledge, the psychological issue of the control problem, which is introduced by J. Zerilli et al., (Zerilli et al.) [54] should be helpful as the precautions for both novices and experts, before we truly know what AI and AI's influence are.

The "control" refers to the human agent's supervisory functions in the human-machine loop, i.e. "both fault diagnosis and management…as well as planning (Zerilli et al. 2019)".[55] Zerilli et al. point out the control problem is caused by "the human agent within a human-machine control loop to become complacent, over-reliant or unduly diffident when faced with the output of a reliable autonomous system", and most importantly, this control problem "… somewhat alarmingly, it seems to afflict experts as much as novices… " (Zerilli et al. 2019).[56] Zerilli et al. analyzed a few reasons that could cause the control problem, e.g. human lack of technical ability or physical limitation to supervise, and one of the reasons is the psychological attitude.

In contrast to the human positive believes in AI is due to human epistemologically misidentified the nature of AI, as I mentioned in previous sections, this psychological attitude of Human is caused by machine's accurate performances. According to J. Zerilli et al., these psychological attitudes of human operator only occur when the automation system is highly reliable, which cause human over-trust the machine and thus change human supervisory behaviors. The highly reliable machine means its less error performance: on the one hand, it makes the human operator become complacency for machine's result and thus won't actively supervise the machine's operation, i.e. the automation complacency; on the other hand, it also makes the human operator inclines to ignore every other information, even include their own senses, i.e. the automation bias[57] (Zerilli et al. 2019). The behavior consequence of these two psychological attitudes should be the alerts while Human interacts with the AI.

As mentioned above, I believe experiences in certain aspects could cause the Human doesn't take the AI's decision for consideration, and thus won't cause any problem as this paper previous mentioned, when providing with AI's decision. But whether this claim is also applicable to the experts when they are facing specific professional decision-making and provide with AI's decision, e.g. when judges to decide pretrial detention, when physicians to diagnose disease and provide medical treatment, when police officer to distribute the police force in certain area…etc., I think we might need more information to determine. And before we truly know what AI is, we should always keep ourselves actively involve and look out all the information when we interact with AI, both the layperson and the expert.

[54] Zerilli, John, Alistair Knott, James Maciaurin, and Colin Gavaghan. "Algorithmic Decision-Making and the Control Problem". Minds and Machines 29(December 2019): 555-578. https://doi.org/10.1007/s11023-019-09513-7

[55] Zerilli et al. 2019 p.559.

[56] Zerilli et al. 2019 p.556.

[57] Zerilli et al. 2019 p.561.

Conclusion

Epistemologically speaking, Artificial Intelligence and the general Technology are heterogeneous for AI's judging criteria lack of the epistemic certainty. While the definitions of AI are still opaque, Humans are paying more attention to the possible advantages than the possible harms that AI could cause. Human in the loop can be an ideal solution to the solely automated decision-making process as the GDPR Art.22 requests, but if we can't recognize the differences between AI and the general technology correctly, human intervention won't work as it meant to be. When we dedicate ourselves to developing and improving the AI, we should ask ourselves as well: What do we Humans really want from the AI?

Reference

ARM and Northstar. "AI today, AI tomorrow: Awareness, acceptance, and anticipation of AI: A global consumer perspective." 2017. http://pages.arm.com/rs/312-SAX-488/images/arm-ai-survey-report.pdf

Article 29 Data Protection Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01). 2018.

Babuta, Alexander. "Innocent Until Predicted Guilty? Artificial Intelligence and Police Decision-Making." *RUSI Newsbrief* Vol. 38, No. 2(March 2018). https://rusi.org/sites/default/files/ 20180329_rusi_newsbrief_vol.38_no.2_babuta_web.pdf

Board of Supervisors, City and County of San Francisco. Administrative Code-Acquisition of Surveillance Ordinance. Accessed February 9 2020. https://sfgov.legistar.com/Legislation Detail.aspx?ID=3953862&GUID=926469C0-A7BA-47D3-BB32-05C2C6D8EB2B

Bristows. Artificial Intelligence: Public Perception, Attitude and Trust. 2018. P.7. https://d1pvkxkakgv4jo cloudfront.net/app/uploads/2019/06/11090555/Artificial-Intelligence-Public-Perception-Attitude-and-Trust.pdf

"Causal Explanation as a Partial Solution to Algorithmic Harms." (paper presented at The 7th Academia Sinica Conference on Law, Science and Technology: Emerging Legal Issues for Artificial Intelligence: Legal Liability, Discrimination, Intellectual Property Rights and Beyond, Taipei, Taiwan, November 26-27, 2018.

Cowgill, Bo. "The Impact of Algorithms on Judical Discretion: Evidence from Regression Discontinuities". (Working Paper)(December 5, 2018). http://www.columbia.edu/~bc2656/papers/Recid Algo.pdf

Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women." *Reuters*, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology*: General, Vol. 144, Issue 1 (February 2015): 114-126. https://doi.org/10.1037/xge0000033

Elsevier. "Artificial Intelligence: How knowledge is created, transferred, and used." January 2019. https://www.elsevier.com/research-intelligence/resource-library/ai-report?utm_source=AI-EC

Entrata. "Artificial Intelligence and Apartment Living: Survey Studies Consumer's Knowledge of and Attitude Toward AI (Report)" and " What Consumers Really Think About AI (post)". August 2019. http://info.entrata.com/newsletters/case_studies/AI/SurveySummary.pdf

Evans, Ian. "Nobody agrees on what AI is"- How Elsevier's report used AI to define the undefinable." *Elsevier*. January 18, 2019. Accessed February 8, 2019. https://www.elsevier.com/connect/nobody-agrees-on-what-ai-is-how-elseviers-report-used-ai-to-define-the-undefinable

GENESYS. "70% of U.S. Employees Hold Positive View of Artificial Intelligence in the Workplace Today. " July 10, 2019. Accessed February 8 2020. https://www.prnewswire.com/news-releases/70-

of-us-employees-hold-positive-view-of-artificial-intelligence-in-the-workplace-today-300882125.html

Gettier, Edmund L. "Is Justified True Belief Knowledge?" *Analysis*, Vol. 23, Issue 6 (June 1963): 121–123. https://doi.org/10.1093/analys/23.6.121

IBM. "IBM to release world's largest annotation dataset for studying bias in facial analysis." Accessed February 8, 2020. https://www.ibm.com/blogs/research/2018/06/ai-facial-analytics/

Lai, Vivian & Chenhao Tan. "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection." In FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 29-38. United States, New York: Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287590.

Logg, J. M., J. A. Minson and D. A. Moore. "Algorithm appreciation: people prefer algorithmic to human judgment." *Organizational Behavior and Human Decision Processes*, Vol.: 151 (February 5, 2019): 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

Marcus, Gary. "Deep Learning: A Critical Appraisal." *ArXiv* abs/1801.00631 (January 2018).

National Institute of Standards and Technology, U.S. Department of Commerce. "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects." (December 19, 2019) P.7. https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

Niiler, Eric. "Can AI Be a Fair Judge in Court? Estonia Thinks So." *WIRED*. March 25, 2019. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/

Northeastern University and Gallup Inc.. "Optimism and Anxiety: Views on the Impact of Artificial Intelligence and Higher Education's Response." October22, 2018. https://perma.cc/57NW-XCQN

Pegasystem."What Consumers Really Think About AI: A global Study." June 19, 2017. https://www.pega.com/insights/resources/what-consumers-really-think-ai-infographic

Ramos, Gretchen A. and Darren Abernethy. "Additional U.S. State Advance the State Privacy Legislation Trend in 2020." *The National Law Review*. January 27,2020. https://www.natlawreview.com/article/additional-us-states-advance-state-privacy-legislation-trend-2020

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Schwab, Klaus. "The Fourth Industrial Revolution: What It Means and How to Respond. Foreign Affairs." *Foreign Affairs*. December 12 2015. Accessed February 8, 2020. https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution

Simonite, Tom. "Machines Taught by Photos Learn a Sexist View of Women." *WIRED*, August 21, 2017. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

Stevenson, Megan T.. "Assessing Risk Assessment in Action." *103 Minnesota Law Review* (2018): 303-384. http://dx.doi.org/10.2139/ssrn.3016088

U.S. Food & Drug Administration. "Software as a Medical Device (SaMD)." Accessed February 8, 2020. https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd

Vaccaro, Michelle and Jim Waldo. "The Effects of Mixing Machine Learning and Human Judgment." *Communications of the ACM* Vol. 62, No.11 (October, 2019): 104 -110. https://doi.org/10.1145/3359338.

Zerilli, John, Alistair Knott, James Maciaurin, and Colin Gavaghan. "Algorithmic Decision-Making and the Control Problem". *Minds and Machines* 29 (December 2019): 555-578. https://doi.org/10.1007/s11023-019-09513-7

ZHIMA Credit, Ant Financial Services Group. Accessed February 8, 2020. https://www.xin.xin/#/home

Author information

Chang-Yun Ku

Academia Sinica, Taiwan

https://www.citi.sinica.edu.tw/pages/evelynku/index_zh.html

Information Law Center, Institutum Iurisprudentiae, Academia Sinica

https://infolaw.iias.sinica.edu.tw/?page_id=562

Research Center for Information Technology Innovation, Academia Sinica

https://www.citi.sinica.edu.tw/people/postdoctoral-fellows

Bridging Natural Language Processing AI techniques and Corporate Communications: towards an integrative model[1]

Dániel Gergő Pintér – Péter Lajos Ihász

Abstract

Today's communication channels and media platforms generate a huge amount of data, which - through advanced AI- (Machine Learning) based techniques - can be leveraged to significantly enhance business networking, improve the efficiency of public relations, management, and extend the possible application areas of communication components. As a sub-discipline of AI, Natural Language Processing (NLP) is frequently utilized in the field of corporate communications (CC) to boost target-group satisfaction through information retrieval and automated dialogue services. This paper gives an overview of the use of NLP in different disciplines of CC, discusses general corporational/organizational practices, and identifies promising research topics for the future while pointing out the ethical aspects of user-data handling and customer engagement. The findings of this synthesizing study are based on primer qualitative research building on the methodology of deep interviews and focus group research involving experts practicing in the fields of CC and NLP. Based on the feedbacks of the participants, a refined CC model was developed, as well as a model mapping conventional NLP techniques onto CC disciplines and tasks they are utilized for.

Keywords: *business management, natural language processing, public relations, corporate communication, deep learning, information society, artificial intelligence, AI ethics*

1. Introduction

1.1. The era of Information Society

With evolving into an information society, access to the social resources and information has been completely rearranged, changing significantly the technological, economic and cultural aspects of everyday life. (Beniger 1986) While in the industrial era the devices and natural resources were definitive for the economy, nowadays knowledge is considered

[1] The authors are grateful to the participants of Budapest Workshop on Philosophy of Technology 2019, and two anonymous reviewers for their valuable critical remarks upon the first version of the present paper; and to the IT and Communication experts for their participation in the deep interviews and focus group research. They are also indebted to Budapest University of Technology and Economics, Ritsumeikan University, Budapest Metropolitan University, Institute for Computer Science and Control (SZTAKI), Media 2.0 Communications and Hungarian Public Relations Association (MPRSZ) for providing a supportive environment for their research.

as the most valuable product. (Castells 1996, 1998) Labour aimed to process information became more significant than direct physical work: the creation, distribution, and manipulation of information is currently the most significant economic and cultural activity, profoundly changing all aspects of social organization. (Pintér 2016; Castells 1997) Thus, digital experience, network-based interaction and unlimited communication have become a basic experience and daily need. (Pintér 2016; Castells 1996) Accordingly, business management started to rely on automated data mining, data analysis, and automated response generation in order to harvest this novel and profound resource. Companies collect and store a large amount of customer data in order to enable better business decisions and gain an advantage in the global market through performing communication that builds on a better understanding of customer needs. (Castells 1997)

1.2. Problem identification, goal and structure

As one pillar of corporate business management, corporate communications - activities aimed to establish and maintain favorable internal and external reputation of the corporation (Riel 2002) - processes and responds to the input of stakeholders and various target audience groups. Many e-commerce websites, for example, allow customers to express their opinions about the products and services the company offers. The reviews are considered not only by fellow customers but with the right information retrieval techniques, these easily obtainable feedbacks serve as a valuable source of information for the companies as well. As another source of information, social media can also be harvested through Artificial Intelligence (AI) - based information retrieval. (Russel and Norvig 2003) Sentiment analysis (SA), for example, is a way to extract semantic information from feedbacks, where opinions, sentiments, emotions, attitudes toward entities and their attributes are computationally identified. Topic extraction, dialogue act classification or summarization are further examples from a wide palette of information retrieval practices. (Carenini et al. 2011; Goldberg 2016)

Extracting customer intelligence from such user-generated content, however, is a challenging task, as it involves dealing with data requiring natural language processing (NLP) techniques. Nevertheless, various machine learning-based NLP methods exist, making possible the effective extraction of customer intelligence, and thus, indirectly enhancing business networking, improving the efficiency of public relations management, and extending the possible application areas of communication components. (Jozefowicz et al. 2016; Goodfellow et al. 2016)

The goal of this study is to underline the seldom researched (and at a first glance non-trivial) connection between the field of communications and computational intelligence. This paper gives an overview on the use of AI-based NLP information retrieval and answer generating practices (NLP techniques) in different disciplines of corporate communications, discusses general practices in a corporational/organizational environment, identifies promising research topics for the future and elaborates on the ethical aspects of AI-based user-data handling and automated communication. As a result of the presented synthesizing study, two models have been developed 1) a model refining the definitions on the disciplines and tasks of CC, and 2) a model mapping NLP techniques onto CC tasks they are applied for. We believe our findings can be utilized by the experts of both fields not only on a theoretical, but on a practical level as well, and that our research can motivate further discussion.

The paper is organized as follows. Section 2 introduces the methodology, elaborating on the details of the experiments for validation. Section 3 describes and specifies the disciplines, tasks and goals of corporate communications and synthesizes between contradicting definitions, introducing our refined model for CC. Section 4 elaborates on the role of AI within today's information society, and specifies the mainstream NLP techniques applicable in CC. In section 5, a model developed to bridge NLP techniques and CC tasks is introduced, along the results of the qualitative research it builds upon. Section 6 exemplifies the application of NLP techniques along the three CC disciplines and discusses the ethical, legal and moral considerations elaborated before. Finally, section 7 concludes the main findings of the study, highlights its limitations, and outlines possible future work.

2. Methodology and experimental setup

The conceptual basis of the study has been laid down through:

1. reviewing the basic literature on CC;
2. synthesizing the fundamental academic definitions of the CC disciplines (Management Communication, Organizational Communication and Marketing Communication);
3. analyzing and redefining one of the most accepted integrative model of CC from a task-focused perspective, differentiating between strategic and operational, as well as internal and external dimensions;
4. summarizing AI-based NLP techniques relevant to CC disciplines based on literature review.

As a primer qualitative research to

- validate and refine our literature review-based findings and conclusions,
- support the reliability of the refined CC model,
- construct a novel NLP-CC model – a model, mapping NLP techniques onto the tasks of the refined CC model – helping to organize the tasks of several general practices used in corporational/organizational environment,
- and to complement our findings with AI ethical and moral aspects,

we conducted:

1. structured deep interviews with communications and AI experts separately. The interviews were 45-60 minutes long one-on-one sessions with 6-6 experts from each field. The questions concentrated on a) what tasks and professional challenges emerge during the daily practice in the fields of CC/NLP, b) what are the possible solutions and best practices in terms of used techniques and services, and c) in which task do they feel that there is a need for improvement (see the questions in detail in

Section 5). The interviews were standardized to have 8 identical questions towards each interviewee so comparisons can be made with confidence between sample subgroups or between different interview periods. The interviewees were selected based on their professional background and seniority level: having a minimum 3 years experience in communications / AI (preferably working with NLP), working mainly on tasks of the operational dimension (e.g. coding, content creation, social media management etc.)

2. focus group research between a heterogeneous group of communication and NLP experts (7-7 people from each field, different from the participants of the deep interviews). The participants were asked about their perceptions, opinions, and attitudes towards a preliminary classification mapping NLP techniques on CC tasks based on the results of the deep interviews. The session was 90 minutes long. Questions were asked in an interactive group setting where participants were free to talk with other group members about the possible discrepancies and insufficiencies of our approach. The authors were present as mere mediators, motivating debate and supporting the group to reach consensus. The participants were selected along the same criteria used in the case of the deep interviews, with the addition that we paid special attention to have at least one expert from each disciplines who works on tasks from the strategic dimension (e.g. reputation management, campaign planning, technical leading, project management etc.), has a minimum of 5 years working experience, and has at least moderate insight to the other field.

Thus, as a result of the deep-interviews, we refined the CC model and built an initial version of the NLP-CC model. The models were further refined in the light of the discussion in the focus group research. The finalized model is illustrated and introduced in Section 5.

3. Corporate Communications as a highly interdisciplinary profession

3.1 Literature review: towards an exact definition

In the scientific literature corporate communication is defined as a set of activities and professional techniques involved in planning, managing and orchestrating all internal and external communications, aimed at creating favourable reputation among stakeholders, target groups and business partners on which the company depends (Fombrun and Riel 2007). According to the definition of Goldhaber (1993, 15), corporate communication is the process of "creating and exchanging messages within a network of interdependent relationships to cope with environmental uncertainty". In his research, Riel (2003, 53) summarizes that CC "is the orchestration of all the instruments in the field of organizational identity (communications, symbols and behaviours of organizational members) in such an attractive and realistic manner as to create or maintain a positive reputation for groups with which the organization has an interdependent relationship".

Frandsen and Johansen (2013) have synthesized the common features of some of the prevailing definitions as seen below:

1. CC is a strategic management function that takes a strategic approach to communication activities and is tied the overall strategy of the company.
2. It integrates external and internal communication activities spread among a series of organizational practices to build, maintain, change and/or repair one or more positive images and/or reputations.
3. All these activities take place inside the relationship with the external and internal stakeholders of the company. (Frandsen and Johansen 2013)

To sum up this interdisciplinary approach, the biggest goal of this profession is twofold: on one hand, to help organizations explain their mission, unique selling proposition and social responsibility; and on the other, to combine the vision of the company, business philosophy and its values into a coherent and credible message to stakeholders and larger audiences, such as employees, media, channel partners and the general public. (Goodman and Hirsch 2010) As Christensen and Cornelissen (2011) pointed out, CC encompasses and coordinates all company's communication activities as an integrated whole with the aim of building and maintaining a valuable image across different stakeholder groups, markets and audiences (Christensen and Cornelissen 2011; Cornelissen 2008).

As it can be seen CC has several, often contradicting or only partly-overlapping definitions, making the selection process of a widely-accepted definition with well formulated specifications to build upon highly difficult and subjective. Thus, in order to achieve the primary goal of this paper - defining CC tasks mappable to NLP techniques - we turned to a discipline-segmenting, task-oriented definition instead of deducing operational communication tasks from a normative description selected from above.

3.2 Corporate communications defined through its sub-disciplines

According to Mazzei's research, CC paradigm prefers the sender's, in this case, the company's perspective, assuming for itself an orchestration role, and justifying centralized control of the entire communication function (Christensen and Cornelissen 2011). This is in accordance with Riel's (1995) well-known, task-oriented definition, which states that CC includes three categories of communication defined by the senders of the communication:

1. Management communication implemented by CEOs, executive managers and teamleaders for: developing offline and online CC strategies; planning and implementing internal and external communications directives; systematically organizing the flow of information; supervising, coordinating, disseminating, revising and monitoring of all the formal channels of communication. Its biggest goals are to develop a shared vision within and besides the organization, gain reputation and maintain trust, enable and manage change processes, support the financial stability and development of the corporation and help employees to grow professionally.

2. Organizational communication includes heterogeneous communication activities: public relations, public affairs, investor relations, government relations, employer branding, corporate social responsibility, investor relations, communication with the labour market, environmental communications, corporate advertising and internal communications.

3. Marketing communication gets the bulk of the budgets in most organizations, encompasses commercial and business communication activities developed to support the sale of goods and services. It typically includes: product advertising, personal selling, promotion, direct marketing, branding and sponsorship activities. (Riel 1995)

This discipline-differentiating definition could be seen as an integrative model linking stakeholders and various internal - and external - target audience groups with the communicational tasks of the organization. This approach gives us a detailed view of CC, which can be used as the basis to link it with NLP techniques.

3.3 A refined CC model

Nevertheless, the tasks and goals mentioned in the description of the disciplines by Riel (1995), partially negligate the synthesization of the common features of CC definitions by Frandsen and Johansen (see Section 3.2). To resolve this discrepancy we redefine management communication as management OF communication (ManC), since it stands for monitoring and analysis-based planning process rather than for the conduction of communication processes in practice. In further accordance with Frandsen and Johansen (2013), we altered Riel's original model from a function-based perspective differentiating between strategic and operational, as well as internal and external (within operational) dimensions. Some tasks were re-specified within the disciplines of Organizational Communication (OC) and Marketing Communication (MC) accounting for the possible intersections. Furthermore, we complemented the model with crisis communication as well, as a crucial but neglected task of CC. (An and Cheng 2010; Pintér 2018) According to the communication scholar Timothy Coombs, as a sub-discipline of public relations profession, crisis communication is designed to protect and defend an organization facing an often unpredictable public challenge to its reputation and performance. (Bundy et al. 2016; Coombs 2007)

The re-definition of the model and re-specification of the tasks were verified and complemented by the valuable remarks of CC experts during the deep interviews and the focus group research (see Section 2 and Table 2). Figure 1 below illustrates the revised model of Riel's.

4. AI in corporate communications

In this section NLP within AI is discussed, and conventional NLP techniques are listed, which can be then mapped onto the tasks of the refined CC model (see Section 3.3).

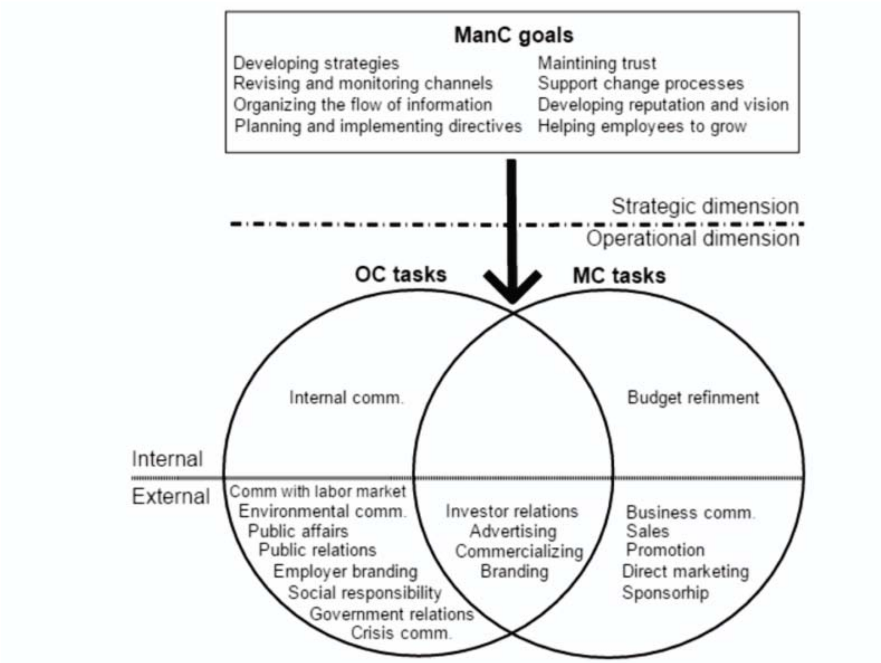


Figure 1: Revised model of CC tasks and disciplines

4.1 Defining NLP within AI

Artificial intelligence (AI) is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. The use of AI conventionally boils down to machine learning (ML), the implementation of algorithm based programs, that perceive their environment and take actions to achieve their goals, while learning from experience to maximize their chance of success. Typical AI / ML algorithms mimic "cognitive" functions, such as classification, pattern recognition, prediction etc. (Poole 1998)

Natural language processing (NLP) is a subfield of AI concerned with the interactions between computer and human, in particular how to develop computational programs (typically MLs) to process, analyze and/or respond to natural language inputs. The three main subfields of natural language processing is automatic speech recognition and synthesization (ASR), natural language understanding (NLU), and natural language generation (NLG). As ASR is related to the pre-processing of NLU and post-processing of NLG, it is often considered as part of them and not as an individual subfield. (Goldberg 2017)

4.2 NLP in the light of CC tasks

Several of the aforementioned CC tasks rely on the summarization and analysis (in alignment with business goals) of data gathered from the target group in order to define and revise strategic steps. Other tasks require the conduction of customer targeting communi-

Table 1. Conventional NLP techniques

Process	Techniques	Description
Analysis	Intent analysis	Inferring the intents of the locutor in the form of dialog acts usually from text source. (Ihász et al. 2019)
	Sentiment analysis/ emotion recognition	Inferring basic emotions (sadness, joy etc.) or sentiments (grouping emotions along their valence into negative, positive and neutral polarities) from text, audio and/or visual sources. Processing multiple source channels in parallel is conventional for this technique. (Ihász et al. 2019; Liu 2015)
	Topic extraction	Inferring one or multiple topics of given text units. (Bun et al. 2002)
	Entity recognition	Extracting target entities (e.g. objects) or named entities (e.g. actual or legal persons) from a given textual source. Extraction from audio source is also possible but unconventional. (Nadeau and Satoshi 2007)
	Content summarization	Summarizing text content into shorter units. (Mani 1999)
Generation	Response generation	Generating responses in the light of user inputs based on hand-made templates or probability-based, deep learning-driven language models. Response generation is achieved by chatbots and preluded by some sort of meaning inference (analysis techniques), and rule-based reasoning. It often involves communication with external information sources (database, ontology or question-answering system). The generated answers are always in text form which can be further synthesized into audio form. (Sordoni et al. 2015)
	Question-Answering	Systems that retrieve information (based on online, web-based or offline, database search) in accordance with the user input and output them in text or audio form. (Xiong et al. 2016)
	Machine translation	Systems translating a source language into target language based on deep learning-driven language models. Input and output is possible both in text and audio.(Kohen 2009)

cation. This is where NLP techniques can boost CC tasks through quick, automatized solutions, able to process complex and robust data. Accordingly, NLP can be subdivided into content analyzing (former NLU and ASR) and content generating (former NLG and ASR) sub-processes.

Table 1 specifies the typical content analysis and content generation techniques along their brief descriptions. The contents were selected as a result of discussion with both AI and CC experts from participating in the deep interviews and focus group research (see Section 2 and Table 2). It is important to note as well, that from a linguistic perspective, the NLP techniques specified in the table are processing semantic - and pragmatic-level (in other words meaning-level) information, preluded by several pre-processing techniques, handling phonetics and phonology - (e.g transcription), morphology - (e.g. part-of-speech-tagging), and syntactic - (e.g. syntactic parsing) level information. In this paper we limit ourselves to discuss only NLP techniques processing meaning-level data, as techniques for lower dimensional information processing could not be effectively associated with CC tasks.

5. Results: the proposed NLP-CC model

In this section the proposed model is presented, mapping NLP techniques onto CC tasks. The mapping is based on the results of the deep interviews and refined through a focus group research.

Figure 2 illustrates an extended CC model, specifying NLP techniques related to CC tasks and goals (and excluding all unrelated tasks and goals). No NLP techniques to the goals of ManC were mapped, since they represent overarching-goals influencing the whole of the company, rather than specific tasks included in OC and MC. These goals are satisfied through the realization of the tasks of the operational-level disciplines (as indicated by the blue arrows). Thus, ManC is, in reality, uses all NLP methods, only in an indirect way, through the result of the implemented operational-level CC tasks. Such tasks (of OC and MC) then are governed accordingly by the strategic-level communication goals, constituting a highly interactive relationship between ManC and OC; and ManC and MC separately. It is important to note, that although ManC achieves a bi-directional information exchange with MC and OC, the two operational-level disciplines - even though sharing common tasks - are not interacting with each other.

The initials below the specific CC tasks indicate the NLP techniques they involve, while the colour (of the initials and the arrows), indicates the direction of the information-flow. Although NLP techniques for content analyzis and generation were differentiated in Table 1, this specification not necessarily overlap the differentiation of information gathering and outward communication. Machine translation for example, which is a content generating technique, is used for content generation and information gathering as well in several OC and MC tasks (indicated by the duplicated use of the initial M in both read and blue colours). The CC tasks which share resonating goals were grouped (indicated by brackets) since they utilize the same set of NLP techniques.

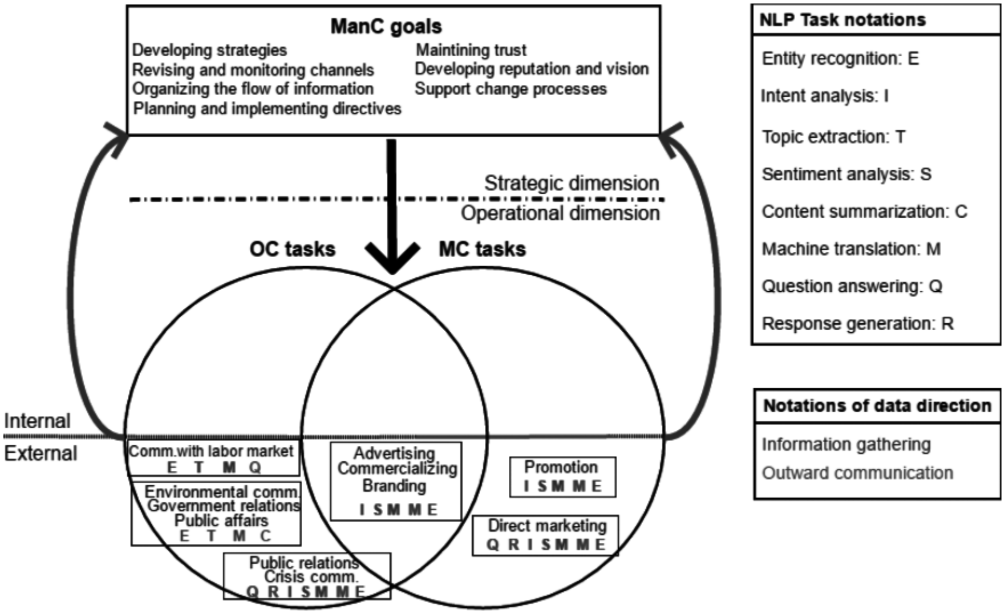


Figure 2: The NLP-CC Model, mapping NLP techniques on CC tasks

Table 2. below specifies the questions and answers of the deep interviews and focus group research the above model was built upon. Ratio of consensus is also indicated, measuring the extent of alignment among a) the 6-6 CC and NLP experts (separated by expertise), participating in the deep interviews, and b) between the 7-7 CC and NLP experts attended the focus group research. The authors concentrated on analyzing the intersection related aspects of CC and NLP as comprehensive as possible, thus even when opinions were highly divided (50% of consensus), they tended to acknowledge the existence of a given problem/aspect. The debate generating questions during the focus group research was mainly identical to the ones asked in the deep interviews (and thus not specified in the table), with a greater emphasis on the questions where the ratio of consensus was initially low, e.g. on the ethical aspects regarding NLP in internal communication. Considering the length limitations of the paper not all concerns were specified in the disagreement section, only one example was provided per question for each group of experts reflecting the most concerning issues. (Over-specifying the separate concerns would shift the balance of the paper which would no longer concentrate on giving a holistic overview on the overlap of NLP and CC.) The results were used to validate and revise the refined CC model, and to support and expanding it to the NLP-CC model.

Table 2. Experimental results: summary of questions and answers

Question	Answer with consent of majority	Ratio of consensus	Main points of disagreement
1. Are the tasks listed (in the refined CC model / Table 1 of NLP techniques) representing your field of expertise to a satisfactory-level?	Yes.	CC experts of deep interviews: 66% (4/6)	Some experts argued that the environmental comm., public affairs and business comm. tasks are fused in some cases. Thus their separation into different tasks and to different disciplines can be inadequate. On the other hand, one expert claimed that the advertising and commercializing tasks are purely MC tasks, and should not be positioned in the intersection of OC and MC.
		NLP experts of deep interviews: 100% (6/6)	All NLP experts agreed on the techniques of Table 1 as conventional NLP techniques.
		Focus group participants:79% (In agreement: 11/14; 6/7 NLP, 5/7 CC experts)	Some CC and NLP experts found the differentiation between the goals of strategic dimension and the tasks of operational dimension ambiguous. CC experts pointed out that the strategic goals and operational tasks are sharing an interactive connection, where the borderline between goal formulation and task implementation is not distinguishable. One NLP expert found it difficult to visualize how the implementation of specific tasks serve several strategic goals in parallel. (E.g. employer branding serves the goals of maintaining trust, supporting change processes and developing reputation as well).

2. What are the CC tasks that are plausible to be solved by any NLP solution? (What CC tasks to keep in the model?)	See the tasks specified in the OC and MC sets of Figure 2.	CC experts of deep interviews: 83% (5/6)	The definition and application of sentiment analysis was debated. (E.g. how it is realized in the commercializing task and is it required at all.)
		NLP experts of deep interviews: 66% (4/6)	Differentiation between activities related to the tasks of MC were debated. (E.g. promotion vs. direct marketing)
		Focus group participants: 57% (In agreement: 8/14; 4/7 NLP, 4/7 CC experts)	There is a lack of mutual understanding between the experts of the two fields. CC experts tend to overestimate the possible use-cases of NLP, while NLP experts having difficulties with recognizing the outcomes of several CC tasks. (E.g. NLP techniques are not applicable on the investor relations task. What is the exact purpose and outcome of environmental comm.)
3. What are the specific NLP techniques that are plausible to apply for the given CC tasks?	See the mapped techniques notated by initials in the OC and MC sets of Figure 2. and a detailed discussion in Section 6.1.	CC experts of deep interviews: 50% (3/6)	The role of machine translation and topic extraction was highly controversial in the case of OC.
		NLP experts of deep interviews: 66% (4/6)	The role of entity recognition was ambiguous in the case of MC. (E.g. how it is realized in the task of promotion.)
		Focus group participants: 50% (In agreement: 7/14; 4/7 NLP, 3/7 CC experts)	There was discrepancy in the NLP-technique- specific grouping of the CC tasks. (E.g. although environmental communication, government relations, and public affairs share partially similar goals, according to some of the participants they might not utilize the same NLP techniques: is machine translation and content summarization required for the inherently outward-directed public affairs task?).

4. What are the NLP techniques that serve the purpose of outward communication / information gathering? (Direction of data-flow.)	See the techniques differentiated by colour in the OC and MC sets of Figure 2.	CC experts of deep interviews: 83% (5/6)	NLP techniques of outward communication was debated to be applicable in the case of some OC tasks. (E.g. since misinformation in crisis communication can threaten personal safety and economical stability, automated response generation may deemed to be dangerous to apply.)
		NLP experts of deep interviews: 66% (4/6)	
		Focus group participants: 86% (In agreement: 12/14; 7/7 NLP, 5/7 CC experts)	
5. Are there any NLP techniques plausible for ManC purposes?	NLP techniques are utilized for the goals of ManC only through the results of OC and MC in an indirect way (indicated by the blue arrows in Figure 2.)	CC experts of deep interviews: 100% (6/6)	It was agreed by all experts that in order for the ManC to successfully fulfill its strategic planning role (towards the operational level) it continually requires input from the operational-level tasks of OC and MC disciplines.
		NLP experts of deep interviews: 50% (3/6)	The realization of strategic and operational dimensions in terms of CC tasks was ambiguous. (E.g. how exactly the use of question-answering for the task of direct marketing serves the realization of ManC strategic goals.)
		Focus group participants: 79% (In agreement: 11/14; 4/7 NLP, 7/7 CC experts)	There was a lack of understanding on the interactive structure of CC disciplines from the NLP experts-side. (Similar to the concerns of the CC experts detailed in the focus group disagreement section of Question 1)

6. Are there any NLP techniques applicable for internal (as opposed to external) OC and MC tasks?	No.	CC experts of deep interviews: 83% (5/6)	The experts from both sides agree on that NLP is not utilized widely (to their knowledge) at the moment for internal communication purposes. However, there were suggestions from both sides on technologies applicable (but currently unexploited) for gathering information from and about the employees of the company. The ethical way of using NLP would be automated and anonymized survey and content analysis (through all the analysis-related NLP techniques mentioned in Table 1) to measure the level of employee engagement; to identify topics of interest; to understand the overall perception about the company. The unethical application would be monitoring and distinguishing certain employees based on the content shared on the official platforms of the company. E.g. fishing for employees planning to leave the company / damages the reputation of the company etc. Both NLP-based information gathering methods could provide inputs for internal communications. Some CC experts also suggested the development of chatbot-like personal assistants to navigate the employees through complex internal platforms / to find specific intra-company information in a centralized way.
		NLP experts of deep interviews: 66% (4/6)	
		Focus group participants: 78% (In agreement: 11/14; 6/7 NLP, 5/7 CC experts)	

7. What are the potential use cases of NLP techniques for CC, that would worth to be implemented in the future?	CC experts of deep interviews: Question - answering could be used for swift information gathering from intra-company sources. E.g. measuring employee-satisfaction via the automated intention/sentiment analyzis of free-input feedback forms.	100% (6/6)	None
	NLP experts of deep interviews: monitoring online behaviour through topic extraction in order to measure employee engagement.	50% (6/3)	E.g. monitoring of private search-history (even on company machines) would raise several ethical and legal issues.
	Focus group participants: internal communication is unexploited (from an NLP utilizing perspective) and should be considered to at least the same extent as external communication, because employees are an equally important target group of CC. The NLP techniques which are implemented for external purposes	100% (14/14)	Although there were some debate related to ethical considerations, nevertheless, all participants agreed on that the power of NLP could be exploited on internal communications as well, if the resources and number of employees render it rational.

8. Are there any ethical problems that can emerge in the application of NLP to CC task in your opinion?	Yes. (See a detailed discussion in Section 6.2.)	CC experts of deep interviews: 83% (5/6)	The artificial agents may be confused to human agents during the official interaction between the customers and the company (e.g. help desk chatbots). In the case of communication with artificial agents legal responsibility is not clarified.
		NLP experts of deep interviews: 50% (3/6)	The process of information gathering and analyzis often induces several ethical problems related to private data management. (See Section 6.2.)
		Focus group participants: 100% (14/14)	Although all experts agreed on that there are several potential ethical problems, opinions from the representatives of the two fields differed on the nature of the possible sources what can endanger the customers. CC experts highlighted the harmfulness of misleading customers, while initially failed to recognise what NLP experts stressed, that information gathering and analysis is often conducted without the full consent and awareness of the customers. NLP experts on the other hand concentrates on the pragmatic side of implementation, and consider less the possible harms the misuse of the application of NLP can cause.

6. Discussion

In this section the results of the primer qualitative research (represented in Table 2 and Figure 2) is discussed with general examples of organizational/corporational practices along two axis: results related to the bridging of NLP and CC; and results related to ethical considerations.

6.1 General practices of applying NLP for CC in organizational/corporational environment

The deep interviews and focus group research revealed several practices implemented in the industry in the relation of CC.

Management of Communication:

Although no NLP techniques can be directly associated with the strategic goals of this discipline, the primer qualitative research revealed that several directives support and motivate the application of computational linguistics in order to improve the overall efficiency of the company. Thus, management must account for AI-related planning and implementation:

- integrating the use of AI into the company's workflow and processes;
- recruiting AI specialists;
- organizing trainings focusing on the best practices of AI-leveraging communications;
- supporting digitalization and the adoption of AI business mindset.

According to the participants of the research, the following NLP-applications of OC and MC give input for strategic planning. The applications listed below are often overarching, utilized in several CC tasks.

Organizational Communication:

- Finding a common thread across heterogenous target groups: mainly utilizing the NLP techniques of entity recognition, sentiment analyzis, and machine translation. It is applied along all of the OC tasks.
- Predict media trends, discrepancies, conflicts and identify deception: mainly utilizing the NLP techniques of entity recognition, text summarization, sentiment analyzis, and machine translation. It is applied mostly in the OC tasks of public affairs, public relations, government relations, social responsibility and crisis communication.
- Issue management, tracking rumors and crisis indicators, determining and avoiding PR crises: mainly utilizing the NLP techniques of entity recognition, text summarization, sentiment analyzis and intent analyzis. It is applied mostly in the OC tasks of public affairs, public relations, government relations, and crisis communication.
- Public Relations and impact measurement, metrics-improvement: mainly utilizing the NLP techniques of entity recognition, sentiment analyzis, and topic extraction. It is applied mostly in the OC tasks of public relations, employer branding, crisis communication and communication with the labour market.
- Creating media lists, corporate-related content, writing data-driven stories, and automatically delivering news: mainly utilizing the NLP techniques of text summarization, entity recognition, response generation and question answering. It is applied mostly in the OC tasks of internal communication, public relations, employer branding and crisis communication.

Marketing Communication:

- Creating more accurate buyer personas, identifying target customers: mainly utilizing the NLP techniques of entity recognition, sentiment analyzis, and topic extraction. It is applied mostly in the MC tasks of business communication, sales, sponsorship and direct marketing.

- Building online chatbots and voice assistants, automating customer service: mainly utilizing the NLP techniques of entity recognition, intent analyzis, response generation and question answering. It is applied mostly in the MC tasks of promotion, sales and direct marketing.
- Marketing measurement, market research and analysis, metrics improvement: mainly utilizing the NLP techniques of entity recognition, sentiment analyzis, content summarization and topic extraction. It is applied mostly in the MC tasks of budget refinement, business communication, sales, direct marketing.

6.2 Ethical and moral considerations of using AI-based NLP techniques in CC

Although this sub-section discusses issues form the receiver's perspective (as opposed to the sender's perspective used to specify the tasks of the developed models), to fully account for the concerns of the participants of the deep interviews and focus group research, and to understand the discussed CC task – related NLP techniques in their full scope, we found it important to elaborate on the main ethical issues that can arise from the use of NLP techniques for CC tasks.

The ever-growing scope of Natural Language Processing AI techniques present new opportunities for corporations to mine customer intelligence and to fine-tune / improve their communications accordingly. On the other hand, the utilization of such techniques often result in handling of confidential data, possibly unwanted by the target group, who is not provided with the means for adequate self-management of data-privacy. As it has been pointed out by the participants during the research conducted, the use of NLP for internal communication purposes, for example, offers several unexploited possibilities, but may lead to unwanted and unauthorized monitoring of the employees. Thus, the unprecedented growth of data is creating not only business opportunities but complex ethical problems as well, with no standardized solutions to deal with them. The rapid pace of development in data science is not met with the understanding of the legal, moral, and cultural issues associated with its collection, storage, and reputation management-oriented analysis. (Mittelstadt and Floridi 2016; Larson 2013)

The aforementioned problems stemming from the careless use of AI technology or insufficient knowledge about how to use them properly can be considered as indicators for organizational crises with unpredictable negative effects. It is thus important to map the ethical aspects, accounting for cases where the interests of the customers' could be potentially harmed, in order to prepare for and possibly avoid such crisis. Without doing so, the mere usage of AI techniques (or any device utilizing such techniques) need to be considered as possible threats not only for the individual customers, but in an indirect way, for the reputation and image of the whole company. (Neuman and Pintér 2019).

Based on the academic literature and the experimental results of this study, we inferred that the most widely discussed issue regarding the usage of NLP techniques for CC purposes is that the requirement of transparency during data-handling is often not satisfied. The main ethical issues are briefly listed below focusing on non-transparent information gathering and interaction practices.

Information gathering and analyzis

- Private data management consent is often given without reading / full understanding of the risks and consequences. The user, however, might not want his data to be analyzed. (Saqr 2017) The failure of proper consent-gathering has been widely discussed and acknowledged. (Mantelero 2014)
- In order to use a given online service, consensus to private data management is obligatory. With no alternative means to use a service, the user is often pushed to accept any terms and conditions and forced to give access to sensitive information.
- There is no time limit on the utilization of gathered customer data. Terms of agreement (and utilization purposes) might change with time. (Saqr 2019)

The users should be adequately and completely informed about the purposes of the data gathering: how it is collected, stored, transferred and processed. The customer should also be informed if a third party will acquire access / participate in processing of her/his data.

- Users should be given clear guarantees that data will not be sold or transferred to other legal entities without their proper consent.
- Although it is not widely implemented, NLP could be used effectively for internal communication purposes, e.g. to measure employee satisfaction and survey attitudes towards certain work-related topics. On the other hand, this can easily lead into the monitoring of the employees (private search-history on company machines etc.) which would raise a new dimension of ethical and legal problems: where is the line between spying and surveying?

Interaction and communication

- Customers may confuse the artificial agents to human agents. When they realize they have been interacting with a chatbot they may feel tricked. As a possible result, the level of trust in the brand decreases which has a negative effect on the company's reputation.
- Portraying chatbots with human-like avatars (with names, images, personal info) may give customers the false impression they are interacting with a real person, which can lead to long lasting emotional / financial harm for both sides.
- Users are often not provided with sufficient information to clarify who is the legal person who can be held responsible in the case they are misinformed by an artificial agent.

To account for the issues listed above, specific cases of data-abuse are need to be identified and analyzed, in order to develop novel approaches which could guarantee the users the necessary knowledge and means for a more effective self-management of data-privacy.

The rapid growth of information necessitates the orchestrated efforts of all parties involved in CC to understand the legal, ethical, and cultural problems we are facing. As conventional practices do not help to solve the issues of our modern information society, we need novel, widely adaptable, but rigorous policies, applicable not just for the current problems but for future challenges as well. (Varley-Winter and Shah 2016) Thus, a general framework should be developed, governing and supervising NLP-based data-handling. It is up for research to further investigate this problem.

7. Concluding remarks, limitations and further research

This study contributes to the fields of communications and computational linguistics with three main findings:

1. A function-based model of CC has been developed building on Riel's original concept. The synthesizing model resolves the discrepancies and contradictions with other definitions / daily practice; differentiates between strategic and operational, as well as internal and external dimensions; and re-specifies several tasks accounting for the possible intersections of disciplines.
2. As an extension of our validated CC model, an NLP-CC model has been developed, mapping NLP techniques on CC tasks. General examples of organizational/corporational practices has also been discussed giving context to the mapping.
3. Our research addressed the ethical aspects of utilizing NLP for CC purposes revealing two potential sources where the customers' interests can be harmed:
   a) There are no policies regulating the use of artificial agents, which would protect the customer from misleading them into the false belief that they are interacting with a real person /grant the customer proper channels to hold someone responsible in case being misinformed by an artificial agent.
   b) The customer is often not properly informed about how, where and when her/his personal data is utilized.

The findings of this study is based on the author's professional experience in the field of communication and computational linguistics supported by primer qualitative research building on the methodology of deep interviews and focus group research involving experts from both fields. Based on the feedbacks of the participants we defined possible associations between CC tasks and NLP techniques. (The qualitative research also revealed several important unexploited possibilities. E.g. NLP is not utilized widely for internal communication purposes, however several technologies applicable for gathering information from and about the employees). However, we outlaid only general examples of typical organizational/corporational practices without mapping them to company specific use-cases which would lead to a deeper understanding about the key factors of the intersection of the two disciplines. Further empirical research is needed to validate and strengthen the representativeness and universality of our findings especially in the cases where the ratio of consent within the collected answers was lower. In addition, surveying the best practices of large, multinational companies typically utilizing customer feedback, such as Amazon, Rakuten etc. and communication and media agencies would provide further insights on the overlap of the discussed fields.

Both fields could further benefit from supporting the results of qualitative research (conducted so far) with quantitative research methodology as well: surveying larger, statistically significant, diverse sub-samples (accounting for demographic, experience-wise etc. variability) of target population (NLP/CC experts) based on and extending the questions of the deep interviews and focus group research (see Table 2) through online questionnaires.

This pioneering study tries to synthesize NLP (a field of sciences supported by quantitative data and exact definitions) and CC (a field of social sciences with less exact, ever-changing definitions). Mapping between the two fields is a complex task with no conventional solutions to the knowledge of the authors. Accordingly, during the creation of the proposed models, the authors forced to make several subjective decisions: what aspects to consider during the selection of participants, how to resolve disagreement/ contradiction between the interviewed experts and to what extent should they rely on the expertise of the interviewed participants and previous work.

Our study concentrates rather on finding the correspondence of NLP and CC, than solving the definitional problems and providing a rigorous differentiation between the disciplines of corporate communication. Although our models were developed from an AI perspective, they also account for several discrepancies of the well-known but overly-flexible and obsolete (in the context of modern information society) framework of CC by Riel. This approach, however, is just one of the many possible ones, which can also be thought as a debate-generating initiative. Instead of concentrating on a holistic perspective as this study did, it could be beneficial to analyze specific use-cases between given NLP techniques and CC tasks on a more detailed level: e.g. conducting case-studies how entity recognition can be utilized for crisis communication purposes, etc. Modern, deep learning-based NLP solutions could be, for example, utilized in crisis communication to: predict organizational crises, to analyze the possible outcomes, to classify different crisis categories, to speed up response-generation or to build AI bots assisting crisis communication experts. (Coombs 2007)

This is a practice-driven study where theory is criticized and refined considering the daily routine. As it has been pointed out by Pinter (2019) - paraphrasing Gregory-Miller's (1988) models of "Public understanding of science" - in fields with strong pragmatic outcomes and dependencies on ever-improving technical tools such as communications and computational linguistics, regular revision of theory in the light of an equally valuable consideration factor: practice, is crucial. Furthermore, involving experts in scientific research to keep track of the emerging challenges and the development they necessitate in this rapidly changing, modern information society is of outmost importance.

Bibliography

An, S. K. and I. H. Cheng. "Crisis Communication Research in Public Relations Journals: Tracking Research Trends Over Thirty Years", in *The Handbook of Crisis Communication* edited by Coombs, W. Timothy; Holladay, Sherry J., 65-89. Oxford, UK., Wiley-Blackwell, 2010. http://dx.doi.org/10.1002/9781444314885.ch3

Beniger, J. R. *The Control Revolution: Technological and Economic Origins of the Information Society*, Cambridge, Mass.: Harvard University Press. 1986.

Bun, K. K. and M. Ishizuka. "Topic extraction from news archive using TF* PDF algorithm. "*Proceedings of the Third International Conference on Web Information Systems Engineering*, Institute for Electrical and Electronics Engineer (2002) http://dx.doi.org/10.1109/WISE.2002.1181645

Bundy, J; M. D. Pfarrer; C. E. Short and W. T. Coombs. "Crises and crisis management: Integration, interpretation, and research development", *Journal of Management* vol 43, no 6 (2016): 1661–1692. http://dx.doi.org/10.1177/0149206316680030

Carenini, G.; G. Murray and R. Ng. *Methods for Mining and Summarizing Text Conversations: Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2011. http://dx.doi.org/10.1145/2348283.2348529

Castells, M. *End of Millennium: The Information Age: Economy, Society and Culture* vol 3. Cambridge, Massachusetts; Oxford, UK: Blackwell. 1998. ISBN 978-0-631-22139-5.

Castells, M. *The Power of Identity, The Information Age: Economy, Society and Culture* vol. 2. Cambridge, Massachusetts; Oxford, UK: Blackwell. 1997. ISBN 978-1-4051-0713-6.

Castells, M. *The Rise of the Network Society, The Information Age: Economy, Society and Culture* vol. I. Cambridge, Massachusetts; Oxford, UK: Blackwell. 1996. ISBN 978-0-631-22140-1.

Christensen, L. T. and J. Cornelissen. "Bridging corporate and organizational communication: review, development and a look to the future", *Management Communication Quarterly* vol 25, no 3 (2011): 383-414. https://doi.org/10.1177/0893318910390194

Coombs, W. T. *Ongoing Crisis Communication: Planning, Managing, and Responding*, Los Angeles: Sage. 2007.

Cornelissen, J. *Corporate Communication: A Guide to Theory and Practice*, Sage, London. 2008.

Frandsen, F. and W. Johansen. "Corporate communication" in *The Routledge Handbook of Language and Professional Communication* edited by Bhatia, Vijay; Bremner, Stephen, Routledge, London. 2013.

Goldberg, Y. "A Primer on Neural Network Models for Natural Language Processing", *Journal of Artificial Intelligence Research* vol 57. (2016):345–420. arXiv:1807.10854

Goldberg, Y. "Neural network methods for natural language processing", *Synthesis Lectures on Human Language Technologies* vol 10, no 1 (2017) https://doi.org/10.2200/S00762ED1V01Y201703HLT037

Goldhaber, G. M. *Organizational Communication*, (6th ed.), Brown & Benchman, Madison, WI. 1993. [originally published in 1974]

Goodfellow, I.; Y. Bengio and A. Courville. *Deep Learning*, MIT Press, 2016.

Goodman, M. B. and P. B. Hirsch. *Corporate Communication*, Peter Lang Inc., International Academic Publishers, New York, 2010. ISBN-13: 978-1433106217

Gregory, J. and S. Miller. *Science in public: communication, culture, and credibility*, NewYork: Plenum.1988.

Ihász, P. L.; M. Kovács and V. V. Kryssanov. "Emotion Recognition through Intentional Context." *International Journal of Affective Engineering* vol 18, no 1 (2019):17-25. https://doi.org/10.5057/ijae.IJAE-D-18-00002

Jozefowicz, R.; O. Vinyals; M. Schuster; N. Shazeer and Y. Wu. *Exploring the Limits of Language Modeling*. 2016. arXiv:1602.02410v2

Koehn, P. *Statistical machine translation*, Cambridge University Press. 2009.

Larson, E. "Building trust in the power of "big data" research to serve the public good." *JAMA: the journal of the American Medical Association* vol 309, no 23 (2013): 2443-2444. https://doi.org/10.1001/jama.2013.5914

Bing, L. *Sentiment analysis: Mining opinions, sentiments, and emotions*, (1st ed.) New York, USA.: Cambridge University Press, 2015.

Mani, I. *Advances in automatic text summarization*. MIT Press, Cambridge, United States. 1999.

Mantelero, A. "The future of consumer data protection in the E.U. Re-thinking the 'notice and consent' paradigm in the new era of predictive analytics." *Computer Law & Security Review*, vol 30 (2014): 643–60. https://doi.org/10.1016/j.clsr.2014.09.004

Mittelstadt, B. D. and L. Floridi. "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts." *Science and Engineering Ethics* vol 22, no 2 (2015): 303-41. https://doi.org/10.1007/s11948-015-9652-2

Neuman, P. and D. G. Pintér. "Technology-based critical phenomena: a Borgmannian approach of crisis prediction." in *Essays in Post-Critical Philosophy of Technology* edited by Héder, Mihály; Nádasi, Eszter, Ch.13: 125-134. Vernon Press. 2019. ISBN: 978-1-62273-457-3

Pintér, D. G. "Revealing the connections between Situational Crisis Communication Theory and Framing Theory with their obstacles and opportunities for improvement in connection with media analysis of police news conference on the 2016 Budapest explosion." *Doctoral Dissertation*, Eötvös Loránd University, 167-168. (2019)

Pintér, D. G.. "Various Challenges of Science Communication in Teaching Generation Z: an Urgent Need for Paradigm Shift and Embracing Digital Learning." *Opus et Educatio* vol 3, no. 6 (2016): 1-25. http://dx.doi.org/10.3311/ope.146

Pintér, D. G. „Media Bias and the Role of User Generated Contents in Crisis Management: a Case-Study about the Communication of the Hungarian Police Forces after 2016 Budapest Explosion." *Corvinus Journal of Sociology and Social Policy* vol 9, no 1 (2018): 101-125. https://doi.org/10.14267/CJSSP.2018.1.05

Poole, D.; R. Goebel and A. Mackworth. *Computational intelligence: A Logical Approach*, New York: Oxford University Press. 1998.

Riel, C. B. M. van., "Defining corporate communication" in *Corporate Communication: A Strategic Approach to Building Reputation*, edited by Bronn, Peggy Simcic; Wiig, Roberta. Gyldendal Norsk Forlag, Oslo, (2003):21-40. ISBN 9780415328265.

Riel, C. B. M. van and C. J. Fombrun. *Essentials Of Corporate Communication*, Abingdon & New York: Routledge. 2007. ISBN 9780415328265.

Riel, C. B. M. van. *Principles of Corporate Communication*, Prentice Hall Europe, Hemel Hempstead. 1995.

Russell, S. J. and P. Norvig. *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall. 2003. ISBN 0-13-790395-2

Saqr, M. "Big data and the emerging ethical challenges," *International Journal of Health Sciences* vol 11, no 4 (2017): 1–2.

Sordoni, A; M. Galley; M. Auli; C. Brockett; Y. Ji; M. Mitchell; J. Y. Nie; J. Gao and B. Dolan. "A neural network approach to context-sensitive generation of conversational responses." *Proceedings of NAACL-HLT* (2015) arXiv:1506.06714

Varley-Winter, O. and H. Shah. "The opportunities and ethics of big data: practical priorities for a national Council of Data Ethics," *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences* vol 374, no 2803 (2016) https://doi.org/10.1098/rsta.2016.0116

Xiong, C.; V. Zhong and R. Socher. "Dynamic coattention networks for question answering." *Proceedings of ICLR* (2017) arXiv:1611.01604

Author information

Dániel Gergő Pintér

https://www.linkedin.com/in/dániel-gergő-pintér-phd-01532487/

Péter Lajos Ihász

https://www.linkedin.com/in/peter-ihasz/

“Not Exactly Reading” – The Nature of Reading in the Era of Screen

Krisztina Szabó

Abstract
Digitalisation and technological innovations have confused our traditional theories of reading; key-concepts of literacy (e.g., reading and writing, text and context, comprehension, reception, and interpretation) have become slurred and vexed, including teaching and assessing reading. This confusion resulted in a debate that, among other issues, has provoked the question of whether digital reading can be considered as reading, or it is just a *distraction* from reading. (Coyle 2008; Badulescu 2016)

To decide on this dilemma, I suggest three attributes: *(1) act*, *(2) reading material*, and *(3) device* that can determine the reading. Concerning their relation, the *device* (the third attribute) determines the *reading material* (the second attribute) and the *act* of reading (first attribute). The consideration of the significance of the *device* is in harmony with McLuhan 1967’s ideas about the determining role of medium and technological determinism; however, it is not a necessary presumption of my ideas.

Based on the above three attributes, I claim that digital reading *is* reading, but a special, *extended* version. Digital reading shows various similarities to print reading but also differences as well; however, these latter are not that significant that could validate the exclusion of digital reading from the category of reading or qualify it as a mere distraction. Moreover, applying digital devices for reading besides traditional reading means a *new opportunity* for comprehension and cognitive development, and these are essentials in improving the reading skills of future generations. Engaging children in the complex mental, physical, and sensual experience that reading can give, irrespectively of the type of reading, is the biggest challenge to accomplish in the field of reading in the 21st century.

Keywords: *Digital Literacy, Attributes of Reading, Reading, Text, Reception, Comprehension, Digital Devices, Technological Determinism*

Introduction[1]

The main question of this paper is whether digital reading can be considered as reading or not. The significance of this issue is two-folded: (A) there must be a cause that query the nature and notion of digital reading, and it would be essential for reading research to

[1] I am grateful to the members of the *Measures of Rationality Research Group*, Department of Philosophy and History of Science, Budapest University of Technology and Economics for the feedback and suggestions, especially to my supervisor, István Danka, Ph. D.

know it. (B) The answer has a huge impact on the future of reading, not just on the act and process, but teaching reading and creating educational material as well as on improving educational systems in the long run. Decisions on involving digital devices in everyday teaching and learning practices largely depend on (or should depend on) our knowledge about digital reading, on how we determine digital reading at first base.

To find the right answer and solve the definitional confusion concerning digital reading is a significant step in the field of technology as well. Today technological innovations of reading focus on two conflicting issues: making digital reading similar to *and* different from print reading at the same time. (Bennett 2020; Lamb 2011) On the one hand, it means innovations that try to copy paper-like *reading material* and book-like reading *devices* to imitate the *act* of print reading in a digital environment as much as they can. On the other hand, innovators tend to emphasise the opportunities, benefits and positive effects of going digital and show that reading in the traditional sense is outdated, and it is worth to switch on the screen to keep up with the rapid changes of the 21st century. The problem is that neither the first nor the second endeavouring seem to reach their goals at present – but in my concern, solving the definitional questions of digital reading could help in these processes.

To define digital reading, we need to identify the attributes on which the decision something is considered as reading or not depends. Reviewing the contemporary literature of reading, I conclude that there are three attributes necessarily describe reading: *(1) act, (2) reading material, and (3) device*. I suggest an examination of digital reading according to these three attributes to decide whether digital reading is reading or not. I consider the answer essential in reading research because of the followings: *if* digital reading *is not reading*, then how should we call, what and how should we think about the process of – let us refer to it as – 'digital content consumption' that, due to technological devices, exposes people's everyday life? *If* digital reading *is reading*, then what are the reasons that question its notion and cause as deep theoretical confusion as the threat of separating digital reading from print reading entirely, exclude it from activities commonly considered as reading, moreover label it as an activity that distracts one from reading?

If we put an end to the debate of print vs. digital reading and make a well-established decision on the notion of digital reading, technological innovators would be free from the constant compel of fulfilling the two folded requirements of making digital reading similar to or different from print reading. Then they could start to focus on what matters: to ensure the opportunity and increase the quality of reading with the help of technological innovations. From the viewpoint of educating future generations and raising the level of people's literacy skills word widely, this seems to be a reasonable and preferable aim.

Accordingly, the first part of this paper (*Section 1-3)* discusses the three attributes of reading, demonstrating how they define both print and digital reading and proving that digital reading, in contrast with other opinions (Badulescu 2016; NEA 2007), *is reading*. Then, in the second part (*Section 4-5*), the focus is on the role of technology in 21st-century reading, by showing the determining force of the third reading attribute (*device*), that influences the second (*reading material*), and together with the third (*act*) reading attributes as well. After a summary of the print vs. digital debate, the paper ends with a discussion of the challenges of reading in the *Digital Age*.

The narrow target group of the paper is one of the researchers and teachers who are constantly working on the improvement of teaching reading, creating digital educational material and digital reading platforms. In a broad sense, the paper addresses all researchers and teachers who are interested in the enhancement of teaching reading literacy skills among children.

1. Three Attributes that Define Reading

Imagine a child sitting in an armchair with a book in her hands, staring at the pages, occasionally turning them, and following the lines on the paper with her eyes. Now, what is she doing? How can we define this activity? Probably, we answer that she is reading a story. Since she is smiling, we could estimate that the story has an impact on her thoughts and feelings, she is constructing some meaning from it, and thus she comprehends it somehow.

Now, let us imagine another child, who is not holding a book, but a technological device, a tablet that can display the story on its screen digitally. The child is staring at the screen, occasionally touching it, pressing a button, and following the lines on the screen with her eyes. Question: what is she doing? How can we define this activity – individually and in contrast with the child-with-the-book-case? One can estimate the answer that she is reading a story, too. Since she is smiling, the story has probably an impact on her thoughts and feelings, she is constructing some meaning from it, and thus she comprehends it somehow. According to this, both children are reading, one is doing a print reading, while the other one is performing a digital reading.

The opinions that label the first case (child-with-the-book) as reading are common; however, in the second case (child-with-the-tablet), judgements are divided. For instance, according to the *National Endowment for the Arts (NEA)* study that discusses the issue of reading of American children in 2008, reading digital contents or learning online is "not reading", but „activities that distract one from reading" (Coyle 2008, 3-4). Moreover, in the newest 2018 study (NEA 2018), NEA still holds to this statement. In line with this, a 2016 study states that digitalism will give us a new experience, "which is not exactly »reading«". (Badulescu 2016, 148) Thus, 'digital reading' intrinsically refers to a distracting activity or new experience, and these are in contrast with reading. However, if children who are consuming digital content are doing something that is "not exactly" reading, then what are they *exactly* doing?

There are those – including me – who claim that digital reading *is* reading as well, and we should handle it accordingly in research, surveys, improvements, and educational practices. The supporters of this opinion do not claim that digital reading is entirely similar to print reading, they are aware of the significant peculiarities of digital reading, but claim that the total exclusion and separation of digital reading from traditional reading is not an intelligible consideration. (Coiro and Dobler 2007; Bolter 2001; Cull 2011; Dougherty 2011; Dyson and Kipping 1998) However, if we would like to go beyond the suppositions mentioned above and decide on the notion of digital reading to separate it or, on the contrary, handle it within the category of reading, it is necessary to discover the attributes that have the defining force to determine an activity like reading. Examining the contemporary lit erature on literacy and comprehension theories, I suggest three possible and sufficient attributes in the process of defining reading. These are (a) the *act* of reading, (b) the *material* of reading, and (c) the *device* of reading. Let me demonstrate the significance and necessity of these three attributes with the two cases (child-with-the-book and child-with-the-tablet) presented at the beginning of this section.

In the child-with-the-book case, the *act* is the physical and mental process of receiving the reading material, the *reading material* is the printed text, and the *device* is the book. Without the *act* (reception), the child is just handling the *device* (book) with the *reading material* (text). Without the text, the child cannot do the reception, just handling the book. Besides, if there were no book, there would not be a surface or display for the text; therefore, the reception would not be carried out.

In the second, child-with-the-tablet case, the *act* is the physical and mental process of receiving the reading material, the *reading material* is the digital text or content[2], and the *device* is the tablet. Without the *act* (reception), the child is just handling the *device* (tablet) with the *reading material* (digital text or content). Without the text or content, the child cannot do the reception, just handling the tablet. Besides, if there were no tablet, there would not be a surface or display for the text or content; therefore, the reception would be impossible to carry out.

Now, at first glance, it seems that both activities share all three attributes of reading; thus, we can state that both children are reading. We could add the label 'printed' in the child-with-the-book case; however, reading in the traditional understanding means reading *printed* materials; thus, this additional label seems unnecessary. In contrast, the label 'digital' in the child-with-the-tablet case seems an adequate refinement, since there the reading material is a digital one and displayed on a digital surface – what is a significant difference comparing to traditional reading. Yet, what is this 'significant difference' that makes an additional specifying label necessary and adequate? Let us have a closer look on the first attribute, namely on the *act of reading*. To do this, the question 'what does it mean to read something' – for the sake of example, to read a printed or a digital book – seems to be the right one.

2. The Act of Reading

In the case of the print, we have several probable alternatives that are easily labelled by the verb 'to read a book' such as to open it and read its full *text* with all the appendixes, footnotes, table of contents from the beginning until the very end (when each letter counts). Alternatively, to open a book, and read *certain parts* of the text according to an exact intention (when only certain parts and information count), either, to open a book and read the full *story* (when the story, the message counts as a whole).

Printed and online literature about the definitions of reading varies according to scientific fields and purposes, no matter whether we talk about printed or online sources. Nowadays, when one would like to know something, the easiest and fastest way is to search for it on the Internet. Doing this, one can found almost 4 billion results for the notion 'reading' via Google and 5 million via Google Scholar. To get a sensible grab on these find-

[2] Section 3 will discuss the issue of the notion of content in the case of digital reading.

ings, it is worth to narrow the focus on online dictionaries, as the most basic literature for someone who intends to find the meaning of a certain notion. Let me do this, and present three examples.

The first one is (a) the *Cambridge Dictionary*, which says that reading is "to say the words that are printed or written […]; to understand and give a particular meaning to written information, a statement, a situation, etc. […]; to look at words or symbols and understand what they mean." ("Reading" and "Digital Reading" definitions 2020) Briefly, *to say words, to understand and give meaning to the written, to do something with the written*. In another phrasing: performing an *act* oriented to the printed or the written. In contrast, (b) *Merriam-Webster Dictionary* refers to reading as "to perform the act of reading words: read something". ("Reading" and "Digital Reading" definitions 2020) Thus, performing an *act* oriented to words. The third one is (c) the *Longman Dictionary*, which considers reading as "perceiving a written text in order to understand its contents" ("Reading" and "Digital Reading" definitions 2020) This can be silent or oral, and this latter can be done with or without understanding the content. Briefly, reading is the *act* of perceiving the written.

If we put these three definitions next to each other, we get the following: (a) performing an *act* oriented to the *print* or the *written*; (b) performing an *act* oriented to *words*; (c) the *act* of perceiving the *written*. One can see that they share the keywords of act, print, written and words, and that they refer two out of the suggested three attributes of reading: (1) *act* (act) and (2) *reading material* (print, written and words). Let us go further with the first one and discuss *act* in detail.

According to the previous definitions of the chosen three online dictionaries, *act* refers to
(a) say (words); understand (meaning); give (meaning), and look (at words or symbols);
(b) perform (reading);
(c) perceive (written text).

If we specify the discussion and turn the focus on scientific studies instead of the selected online dictionaries, we face various tenors of defining reading. Including but not limited to, here are four examples that are worth mentioning here. According to them, reading is:

(a) "a number of interactive processes between the reader and the text, in which readers use their knowledge to build, to create, and to construct meaning." ("What is Reading?" 2020)

(b) "a process undertaken to reduce uncertainty about meanings a text conveys. The process results from a negotiation of meaning between the text and its reader." ("Reading" 2020)

(c) "the act of constructing meaning from *text […] The act of reading is supported by reader motivation and positive reader affect.*" (Afflerbach 2017)

(d) "a temporal activity, and one that is not linear". (Iser 1974, 277)

From these definitions, the following keywords are important:
(a) interactive, processes to build, to create, to construct (meaning)
(b) negotiation (of meaning)
(c) the act (of constructing meaning)
(d) not linear activity

The importance of this brief definitional analysis is to demonstrate that reading by nature is not a passive reception, but a complex and constructive mental activity. Readers actively do something with the reading material both in a physical and in a cognitive sense. There are parts of this activity that are explicit, mostly the physical ones (such as saying, performing, looking), while others, the cognitive ones, are hidden (e.g., understanding, perceiving, giving, creating, constructing meaning). These latter were in focus for a long time when debates were about the mental activity of readers. Theories about the reader who is a passive receiver, who do not have any influence on reading, accept the reading material as it is, are exploded notions – especially in the case of digital reading, where the increased mental activity is needed to keep up with the dynamic reading environment. (Snowling and Hulme 2007)

Although these above presented reading definitions did not refer directly to print or digital reading, at this point we do not have any reason to think the contrary, because they say nothing specific about the medium (e.g., book, newspaper, and flyer) or the genre (e.g., high-quality literature, poem, letter, and essay) of reading, but a live mental and physical activity, that creates a connection between the *reader* and the *text* while constructing meaning. But what if the definitions exclude digital reading indeed? Let us have a closer look at the child-with-the-tablet case.

The question and the possible answers seem to be the same: what does it mean to read a book? To open a digital book on an electronic device and read its full text (with all the appendixes, footnotes, table of contents), from the beginning until the very end (when each letter counts)? Alternatively, to open a digital book and read *certain parts* of the text according to an exact intention or read the full *story* of it (when the story and the message do count as a whole).

By conducting the same research in the dictionaries mentioned above for definitions, as in the case of print reading, one can find the following concerning digital reading. (a) *Cambridge Dictionary*: "You can also search for digital or reading." ("Reading" and "Digital Reading" definitions 2020) (b) *Merriam-Webster*: "The word you've entered isn't in the dictionary." ("Reading" and "Digital Reading" definitions 2020) (c) *Longman:* "Did you mean capital gearing; digital native; sight-reading?" ("Reading" and "Digital Reading" definitions 2020) Surprisingly, it seems that (online!) dictionaries do not consider the notion of digital reading. However, searching for the keyword 'digital reading' results in almost 3 billion via *Google* and 4 million via *Google Scholar*. According to these findings, digital reading is

(a) reading digital or electronic text via screen where the bearer of the text is a digital/electronic device.

(b) reading in a digital environment.

(c) online reading. (Coiro and Dobler 2007; Brown 2001; Nicholas and Clark 2012)

The first thing to notice here is that digital reading is reading *by definition*; it is originally described with the notion of reading. The second is that the reading material (digital, electronic, online text) and the reading device and environment (digital/electronic, online) seem to have an important part that it is worth to make a distinction between offline and online digital reading, as follows:

Offline reading is reading digital or electronic texts, which are not connected to the Internet. They are mostly interpretations of printed texts or texts where hyperlinks do not lead readers out of the text. In contrast, *online reading* is reading digital or electronic text, which is connected to the Internet. They are not just interpretations, but interactive texts where hyperlinks lead out readers from the text. “Online reading is the process of extracting meaning from a text that is in a digital format. Also called digital reading”. (Nordquist 2019) Here three issues do need further discussion:

(a) process,
(b) extracting meaning,
(c) text that is in a digital format.

The first one (a) process remained the same act as the one in the printed case in the sense of the task of word recognition, encoding, text-model formation, strategic processing. (b) Extracting meaning is the main purpose of digital reading, and it is also about the act of finding the meaning of the actual text. However, the main change comparing print reading is exactly in the latter, namely in (c) text, because texts in digital format are “still words being taken in on a computer screen” (Gomez 2008, 44), but completed with additional illustrative and explanatory digital elements, that we call together as digital content[3].

Now, according to the suggested three attributes, print and digital reading seems to share the first one: both have the attribute of the *act*. However, the question remains, are these two activities, the *act* of reading a *printed book* and reading a *digital book* the same or not? Are they equal to reading the news from a crinkly newspaper or on the website of The New York Times, for example? Alternatively, to the case of reading a handwritten letter or an e-mail? Do they remain similar if we change the object of the act of reading, namely the *reading material*? What is the reading material *at all*? Now, the next section is going to discuss the issue of *reading material* as the second attribute necessary to define reading.

3. Reading Material

The major question of this section is how to define text or content in the digital age. Here the necessity of distinguishing text and content is rooted in the referential doubtfulness of digital words, symbols, and other visual elements of digital devices, practically of screens. The notion of text is as complex as the notion of reading is and has plenty of definitions according to scientific fields. “The term text has not been easy to define since 1960s. It was first made difficult by the poststructuralist writers, such as Derrida, Barthes, and Foucault, but their notion of the text and their own texts had relatively little impact on the educational community. Now the computer, which is indeed having a great impact on educational theory and practice, has presented further complications.” (Bolter 1998, 3)

In the simplest form, everything is text *what is written*. Alternatively, as the *Organisation for Economic Cooperation and Development (OECD) Programme for International Student Assessments (PISA)*’s official framework documents on literacy says, “the phrase text is meant to include all language as used in its graphic form: handwritten, printed or screen-based.” (PISA2018 2016, 13). The decision whether something is a text or not is normally quite easy in the case of handwriting: we do not trouble much with the question whether a shopping list for our husband, a short note for a colleague, or grandma’s recipe are texts or not. We consider them as texts, while drawn charts, funny pictures on the margin, or il-

[3] I will discuss the issue of content in the next section.

lustrations in an old codex are what they are: additional explanatory or illustrative elements, but not parts of the texts. If we sketch a route to explain to a tourist the shortest way to a building, for instance, we call it a map and not a text. If we write a guide without a sketch, it is a text. If we add a schematic map to the margin, we still consider the whole thing as text. If we remove the drawing, it is still a text. However, if we remove the lines and only the map remains, the nature of the creation changes, and it became a map or a picture again, and it is not a text anymore. If we put this map-example into the printed environment, individual guides without drawings or guides with drawings are considered as texts, while a sketch of a route without phrased instructions is just a map.

It seems that the line between text and non-text lies in the proportion of written words to other visual elements, both in the cases of handwriting and print. If the amount of words is higher than the amount of other visual elements, we talk about a text; while it is smaller, we talk about something else. However, if we would like to define the demarcation of texts and non-texts in a quantitative form, and give the exact proportion, percentage, and numbers, we immediately found ourselves in a moorland. Let me show this by an example: if we go to a copy shop to print something out in colour, and the number of words and the size of the picture is bigger than a given percentage of the paper, we have to pay a higher amount of money. Nevertheless, this "given percentage" changes according to the price list of the actual copy shop. Now, following this string, we can try to define it by giving an exact percentage, such as a text is something in which the amount of visual elements do not go beyond 50% of the entire content. Then we must deal with questions such as what a visual element is, what is whole content, why 50% etc. Do we talk about screen size and the amount of visible text on the screen without scrolling or the complete content (e.g., an article) what we scroll up and down on the screen?

In the specific cases of a mathematical derivation, a formal logic explanation, a picture book, or a comic strip, for instance, the distinction obscure again. If the amount of mathematical formulas is smaller than the word-phrased explanation, we call it as a text, while it is on the contrary, we will not call it as a text anymore – this does not seem to be a rational separation. In another case, when a philosopher is writing an argumentative paper on formal logic, applying the current logical formulas in her paper, and the end, the amount of these formulas are higher than the word-phrased, descriptive explanations, will the paper lost its textual nature? It does not seem quite right. In the example of a comic strip, we normally ask, have you ever read the Batman comics; and we do not ask, have you ever seen (the pictures) of the Batman comics? Briefly, it seems that the two cases of handwritten and printed (in other words, non-digital/non-electronic or written or paper-based) wordings can be considered as equals in the regard of their textual nature.[4] However, when we involve digital (in other words, on-screen/electronic/typed/non-printed/online) text, the issue becomes more complex.

Digital text is "an electronic version of a written text. Digital Text can be found on the Internet or on your computer or on a variety of hand-held electronic devices. […] By nature, digital text is more flexible. It can be searched, rearranged, condensed, annotated, or read aloud by a computer". ("Redefining Literacy" 2020) However, if you have a digital

[4] Here I do not discuss the linguistics approach of textuality and the notion of text, because it would go beyond the scope of the paper.

text to read, you usually face other things besides text, such as visual elements, audios, and videos, built-in interactive tools – briefly: contents. What is a content? “Digital content is any content that exists in the form of digital data. Also known as digital media, digital content is stored on digital or analog [sic!] storage in specific formats. Forms of digital content include information that is digitally broadcast, streamed, or contained in computer files. Viewed narrowly, digital content includes popular media types, while a broader approach considers any digital information (e.g., digitally updated weather forecasts, GPS maps, and so on) as digital content.” (“Digital Content” definition1 2020) Briefly, here we are talking about text, audio and video files, graphics, animations, images, and information available for download from or distribution on electronic media, such as e-books or iTunes songs. Basically, “if you are on the Internet, most likely you are looking at, watching, or listening to a piece of digital content” (“Digital Content” definition2 2020). However, in this sense, if everything on the Internet can be considered as content, and digital reading is consuming content, then watching a movie, listening to an audiobook, playing an online game should be also considered as reading, and this would be apparent nonsense.[5]

Then what can help us to understand and have a grip on the notion of digital text and content? Two things: the distinction of online and digital text and the notion of hypertext. An online text is more than a digitalised version of a printed text because the online nature of it essentially modifies its reading, meaning, and comprehension. Researchers call these “hypertexts”, which are “linked to each other with hyperlinks so we can easily switch and jump between them, like in a kind of eternal, never-ending and always refreshed text.” (Szabó 2015, 171) Hypertexts, also because of the online space, naturally “live together” with visual elements. This connection could be so complex that sometimes it is difficult to decide what is related to the main text, and what is just an additional illustrative or design element or a supporting icon of the digital device. However, this should not be too surprising: if the connection between an offline text and the visual is so strong and complex, then it should be at least the same in the case of online texts, too. Thus, reading material in a digital environment can be digital text or hypertext, and this latter includes additional interactive visual elements, but both types of texts reserve their textual nature. Videos, audios, and games are contents, can be part of digital reading material, but we watch them, listen to them, play them, but do not *read* them. At this point, I agree with the statement of the NEA study: reading digital content is “not reading” (Coyle 2008, 3-4), however, I still contradict that learning online is “not reading”, but an activity „that distract one from reading” (Coyle 2008, 3-4). Here it is important that “learning” means reading, comprehending, and memorising a text, and I exclude activities such as learning skills by playing logical games or learning foreign words by listening to their pronunciations, etc.

Now, the child-with-the-tablet case looks like as follows: she is performing the *act* of reading digital text/hypertext as *reading material*, and this latter could be a digital book or article or interactive storybook as well as an e-mail or a Facebook post. Now, we have successfully determined two out of three attributes of reading; thus, it is time to turn our at-

[5] Here I do not discuss the role of visuality in digital text, because it would go beyond the scope of the paper.

tention to the last one, namely on the *device*. The next section discusses the issue of the *device* and the question of whether it can overwrite the determining force of the other two attributes and exclude digital reading solely from the category of reading or not.

4. Reading Device and Technological Determinism

The huge influence of technological improvements on reading is salient, even from a brief overview of the history of reading. Not just on how and what we read but also the amount and spread of the act of reading. Shifting devices (from stone table to papyrus and parchment, then paper, at first handwritten, later printed), were huge steps that formed and improved writing methods as well as reading material. The more reading device became available, the more people got the chance to learn reading and perform the act of it. As reading devices improved, reading material became complex, and variant and the target group of reading widened. The privileged status of the act of reading has lost, and it became a common thing to do; now, being literate is a fundamental part of modern, educated societies. (Snowling and Hulme 2007; Baron 2009; Fischer 2003)

At first, reading was a social activity, when the one educated member of the community read aloud the written to the audience. Then, as time went by and technology changed, reading was slowly lowered, became silent and individual. People learned to read alone, reading devices such as books and newspapers gained their persistent roles, at first at institutions, then at homes. The availability of reading devices and materials naturally improved teaching reading, thus people's literacy skills. When text went on screen and became available all over the world, the amount of reading material, and the opportunities to read them suddenly grow endless. Today, in the Digital Era, when the dynamic hypertext, the opportunity of immediate feedback, editing, and storyline-forming rule the online space, reading seems to be social again. As *Bob Stein*, creator of the *Institute for the Future of the Book* puts it: „As sure as I was in 1992 that the future of the book was on screens, I'm now sure that it's social […] There's nothing ideal about reading by yourself […] That's just the way we did it for a long time." (Chant 2016) Skimming online content means maintaining "a jumping-off point for further conversations with people around the world". (Chant 2016)

However, due to technological improvements, the key-concepts of literacy as the fundament of communication, cognition, and learning became slurred and vexed. The change from print to digital is not just a simple platform shift as the previous changes, but a cardinal step in the history of reading. Today the question is not about the life and death of printed books but the future of reading. If we agree with McLuhan's theory of technological determinism, every medium shift changes culture since mediums are human perceptions, thus mediums have more power on society than the message itself. (McLuhan 1964) It determines what, how, and when we read, and thanks to the algorithms, do a far better job than human editors do. The consequence of this process is that digital texts are changing according to the requirements of electronic devices, and so are the print texts to keep up with the rapid digital transformation.

This previous issue can be easily observed through the debate about a choice of preference concerning print vs. digital. (Baron 2015) This is a quite heated debate about relatively subjective things, namely what device is comfortable and effective for individual readers. Much research aimed to assess readers' opinions about print vs. digital reading by listing the pros and cons based on mostly on the reading device. Here I will not discuss them in detail but give a summary of some key points from the comparison, as follows.

Print books can be heavy, complicated to carry along, and have physical limitations in size, weight, and content compared to digital books. However, print books are not as hard for the eyes as digital ones; they can be easily noted on the margins, and the experience of touching, holding, flipping, and smelling them help in the reading comprehension process and memory. Digital books are easily shared, accessed, bought, and loaned from digital libraries, and with built-in digital tools such as vocabulary, searching function, and the opportunity of digital annotations is very practicable. However, they are easily injured, vulnerable, and go dead without regular charging, while print books last long but take up much room. Research shows that readers chose texts to the reading device, and not vice versa. Meaning that if a digital reading device (e.g., tablet) is more available than a print device (e.g., book), then the reader will choose according to the device and not the text. This would not be a scandal in the medieval ages when people did the same and read what was available in print, but today, when we have nearly everything in print, it seems at least strange to let our reading choices limited by the reading device. (Baron 2015)

The significance of the previous will be clear when we realise that the debate is about the effects of technological innovation on reading and education as well. I said education since reading and teaching reading is a fundamental educational issue. One could presume that technological innovations are in the service of training better readers; however, this is not that simple. There is much research about the topic of children's digital reading, and it seems that in some cases digital environment does not help; on the contrary, it distracts children and has a bad impact on their reading performance. Are we becoming lazy at reading? Or "studies clearly show that the addition of technology in the classroom actually has a positive effect on our children's reading and writing." (Konnikova 2014) However, other research claims exactly the opposite, namely that "the more we read online, the more likely we were to move quickly, without stopping to ponder any one [sic!] thought." (Konnikova 2014) "Good reading in print doesn't necessarily translate to good reading on-screen." (Coiro and Dobler 2007) Thus, on the one hand, some claim that format matters, and because of technological improvements, children read less, and their reading performance is poor.

On the other hand, some claim that format does not matter, but the storyline and the process of constructing meaning in a complex digital environment. As David Gatward (2017) puts it, it is "foolish to think that children and teenagers don't read when their primary mode of communication is the written word. […] "Do we care about how children and teenagers enjoy reading, or are we more interested in them meeting our ideal of what a reader is […]?" (Gatward 2017) This question seems to be adequate and has great significance in the 21st century – as the next section aims to show.

5. The Reading Challenge of the 21st Century

After discussing the previous issues of reading, one can ask the question of whether innovators, researchers, teachers, and those who are interested in the field of digital reading have any idea what are they doing? Because it seems that we have been collecting materials in electronic format and digitising books without having a vague consensus about the notion of digital reading or at least a clear decision on the question of whether digital reading is reading or not. We have debates on these issues that can lead us to unity; however, at present, discussions seem to freeze rather than improve. While there are actors, such as the researchers of the NEA, who still claim that new technologies are antagonist to reading, and digital reading is not reading, others are forced to stay in the legitimate-digital-reading discussion rather than level up and start a conversation about opportunities of applying new technologies in the service of reading.

However, there are remarkable tenors to improve the activity of reading by technological innovation (e.g., paper-like displays, inbuilt programs that let us create handwritten notes and marks in digital texts, and find solutions for eye-pain caused by looking at screens for too long), but these are not undisputedly successful. Much research shows that despite the genius and available reading tools, people still prefer print to screen. Research shows that students usually print digital learning materials when they prepare for an exam would like to have a better understanding of texts or feel the same experience and joy as they feel while reading a printed book), and in this case, technological innovations still cannot help seemingly. (Baron 2009) Thus, Karen Coyle's (2008) question, "what technology would make the reading of electronic books appealing?" rightfully emerges. As she continues, it would be "important to have at least a tentative answer to this question before we commit fully to the digitization of library resources." (Coyle 2008) At present, two tendencies can be easily detected in the field of reading device innovation: one is about to make digital reading as similar to print reading as they can. The other one is to express differences between the two types of devices and alienate digital reading from print reading as much as they can. If we would like to approach these two inordinate tendencies, at first, we have to understand and reframe our concept of reading, if it is necessary. To reveal the fundamental similarities and differences of print and digital reading not because of the sake of competition or to proclaim the dominance of one in contrast with the other, but to find their right role and place in 21st-century reading, is a crucial step in the long run. In my consideration, accepting that digital reading is reading can be a starting point in this challenge. Finding a reasonable definition for digital reading helps to describe its true nature and understanding and interpreting digital reading can help to understand and interpret print reading in the digital age.

This latter is also important, but not a much-discussed issue. Print reading is an activity as well as digital reading; it also dynamically changes and needs new models. Thus, when we are talking about traditional reading and text, we must face the truth that they are not something unchangeable and constant things, but living notions are varying according to cultures, eras, and technology. Print reading is also changing with the innovations and examine the effects of these technological changes on print reading is essential. I suppose that we can find many attributes that try to fit the actual digital trends and form and fit print reading materials to digital ones. Thus, in the strong movement to make print content, print text to popular and appealing to readers, even the most traditional texts and books should keep up with new trends to reach the same experience as before. However, it is important to keep in mind that the duty of technology "was never the intent to replace the human experiences that are all around us. Rather, it's a tool that enables us to remove friction and frustration in the places where doing so can make the experience more meaningful or convenient." (Bennett 2020)

Therefore, the challenge of reading in the Era of Screen is to accomplish and bring back the missing experience of classic reading: engagement, emotion, and inner motivation, complex mental, physical, and sensual experience that makes print reading special. This is the way that we could eliminate the amount of distraction in screen reading so that it would be an extended version of reading, and not just a poor, ineffective but practical replacement of the old print reading.

Conclusion

This paper focused on the question of whether digital reading is reading or not. To decide on this question, I suggested three attributes that necessarily define reading: *(1) act, (2) reading material, and (3) device*. According to these, I concluded that digital reading is a specific, extended kind of reading that shows similarities with traditional reading, but significant differences as well. However, these latter do not deprive the digital reading of its reading nature or exclude it from the category of reading.

Besides the demonstration of how the three attributes define both print and digital reading, the paper discussed the role of technology in 21st-century reading, by showing the determining force of the third reading attribute (*device*) over the other two attributes. After a summary of the paper vs. screen debate, the paper closed with an outlook on future challenges of reading, emphasising the importance of theoretical clarity in the field of literacy. Reading is an activity; it constantly changes, no matter whether we are talking about print or digital reading. Thus, revising old definitions and creating new ones in accordance with technological improvements and finding a well-established definition of digital reading will ensure the opportunity and increase the quality of literacy skills improvement in the Digital Age. The paper aims to support these tenors with drawing attention on issues rooted in old school reading approaches and may be of interest to those researchers and teachers who are continually working on the improvement of teaching reading, creating digital educational material and digital reading platforms.

References

Afflerbach, Peter. Understanding and Using Reading Assessment, K–12. 3rd Edition. 2017. Accessed February 15, 2020. http://www.ascd.org/publications/books/117050/chapters/Important-Issues-and-Concepts-in-Reading-Assessment.aspx.

Badulescu, Dana. “Reading in the Digital Age”. *Philologica Jassyensia*. 12, no. 1(23) (2016): 139-149.

Baron, Dennis. *A Better Pencil. Readers. Writers, and the Digital Revolution*. New York: Oxford University Press. 2009.

Baron, Naomi. *Words On Screen. The Fate of Reading in a Digital Word.* New York: Oxford University Press. 2015.

Bennett, Linda. „The print versus digital debate rages on, but should there really be a debate at all?” Accessed February 15, 2020. http://www.adrenalineagency.com/blog/print-versus-digital-great-debate/.

Bolter, J. D. “Hypertext and the question of visual literacy”. In D. Reinking, M. C. McKenna, L. D. Labbo, and R. D. Kieffer (Eds.). *Handbook of literacy and technology: Transformations in a post-typographical world.* (pp. 3-13). Mahwah, MJ: Lawrence Erlbaum Associates. 1998.

Bolter, J. D. *Writing space: The computer, hypertext, and the remediation of print*. Second Edition. Mahwah, N.J., London: Lawrence Erlbaum. 2001.

Brown, J. Gary. “Beyond print: reading digitally”. Library Hi Tech, Vol. 19, no. 4 (2001): 390-399. Accessed https://www.emerald.com/insight/content/doi/10.1108/07378830110412456/full/html.

Chant, Ian. “The Future of Reading. Designing the Future”. *Library Journal.* Accessed September 13, 2016. https://www.libraryjournal.com/?detailStory=the-future-of-reading-designing-the-future.

Coiro, Julie, and Elisabeth Dobler. “Exploring the online reading comprehension strategies usedby sixth-grade skilled readers to search for and locate information on the Internet”. *Reading Research Quarterly*. 42, No. 2, (April/May/June 2007): 214–257. International Reading Association. doi:10.1598/RRQ.42.2.2. Accessed https://ila.onlinelibrary.wiley.com/doi/epdf/10.1598/RRQ.42.2.2.

Coyle, Karen. „E-Reading.” *Managing Technology, Journal of Academic Librarianship* 34, no. 2 (March 2008): 160-162. https://kcoyle.net/jal_34_2.html.

Cull, B. W. (2011): Reading revolutions: Online digital text and implications for reading in academe. *First Monday*. Vol. 16. No. 6. http://firstmonday.org/ojs/index.php/fm/article/view/3340/2985. Last access: 29. 02. 2016.

Dougherty, W. C. (2011): The book is dead, long live the book! *Managing Technology*. http://www.sciencedirect.com/science/article/pii/S0099133311001959. Last access: 29. 02. 2016.

Dyson, M. C., Kipping, G. J. (1998): Exploring the effect of layout on reading from screen, in: Hersch, R. D., André J., Brown H. (Eds.) *Electronic publishing, artistic imaging, and digital typography: Seventh International Conference on Electronic Publishing: Proceedings*. Berlin: Springer–Verlag. pp. 294-304.

Ficher, Steven Roger. *A History of Reading*. London: Reaktion Books Ltd. 2003.

Fitzer, Kim R., and James B. Hale. “Reading and the Brain: Strategies for Decoding, Fluency, and Comprehension” Accessed February 15, 2020. https://www.ldatschool.ca/teaching-the-brain-to-read-strategies-for-enhancing-reading-decoding-fluency-and-comprehension/.

Gatward, David. “The book is dead. Long live reading” *Youth and Children’s Work*. March 2017. Accessed February 15, 2020. https://www.youthandchildrens.work/content/view/full/738158.

Gomez, Jeff. *Print is dead: books in our digital age.* New York: St. Martin’s Press. 2008

Halpern, Sue. “Are We Puppets in a Wired World?” The New York Review of Books. November 7, 2013. Accessed https://www.nybooks.com/articles/2013/11/07/are-we-puppets-wired-world/.

Iser, Wolfgang. *The Implied Reader: Patterns of Communication in Prose from Bunyan to Beckett*. Baltimore and London: Johns Hopkins University Press. 1974.

Iyengar, Sunil et al. *U.S. Trends in Arts Attendance and Literary Reading: 2002-2017*. Washington: NEA Office of Research & Analysis, 2018. Accessed February 15, 2020. https://www.arts.gov/sites/default/files/2017-sppapreviewREV-sept2018.pdf.

Konnikova, Maria. “Being a Better Online Reader”. *The New Yorker*. July 16, 2014. Accessed https://www.newyorker.com/science/maria-konnikova/being-a-better-online-reader#.

Kucirkova Natalia, Karen Littleton. “The digital reading habits of children: A National survey of parents’ perceptions of and practices in relation to children’s reading for pleasure with print and digital books, Book Trust. March 2016. Accessed http://www.booktrust.org.uk/news-andblogs/news/1371.

Lamb, Annette. “Reading Redefined for a Transmedia Universe”. *Learning and Leading With Technology*. November 2011. ISTE (International Society for Technology in Education. U.S. & Canada. Accessed https://scholarworks.iupui.edu/bitstream/handle/1805/8636/39-3.pdf?sequence=1.

Mioduser, D, H Tur-Kaspa, and I Leitner. “The Learning Value of Computer-Based Instruction of Early Reading Skills.” Journal of Computer-Assisted Learning 16 (2000): 54-63.

NEA2007. “To Read or Not To Read. A Question of National Consequence”. Research Report #47. National Endowment for the Arts. Washington, DC. Accessed November 2007. https://www.arts.gov/sites/default/files/ToRead.pdf.

NEA2018. “U.S. Trends in Arts Attendance and Literary Reading: 2002-2017. A First Look at Results from the 2017 Survey of Public Participation in the Arts”. National Endowment for the Arts. Washington, DC. Accessed September 2018. https://www.arts.gov/sites/default/files/2017-sppapreviewREV-sept2018.pdf.

Nicholas, David and David Clark. “Reading' in the digital environment”. Learned Publishing, 25 (2012): 93–98. Accessed file:///C:/Users/Philos_hp430_01/Desktop/reading_LP.pdf.

Nordquist, Richard. “Online reading. Glossary of Grammatical and Rhetorical Terms”. *ToughtCo.*Updated March 20, 2019. https://www.thoughtco.com/what-is-online-reading-1691357.

Pearson, David P, Richard E. Ferdig, Robert L. Jr. Blomeyer, and Juan Moran. “The Effects of Technology on Reading Performance in the Middle-School Grades: A Meta-Analysis With Recommendations for Policy.” North Central Regional Education Laboratory, 2005.

Snowling, Margaret J. and Charles Hulme. *The Science of Reading. A Handbook.* UK: Blackwell Publishing. 2007.

Szabó, Krisztina. “Digital and Visual Literacy: The Role of Visuality in Contemporary Online Reading”. In. *In the Beginning was the Image: The Omnipresence of Pictures: Time, Truth, Tradition. Series Visual Learning 6.* Frankfurt am Main: Peter Lang GmbH, Internationaler Verlag der Wissenschaften. 2016. 103-112.

PISA2018 Draft Analytical Frameworks May 2016. Paris: OECD Publication Service. 2016.

“Reading”. Foreign Language Teaching Methods. Accessed February 15, 2020. https://coerll.utexas.edu/methods/modules/reading/01/.

“Redefining Literacy”. *Transforming Literacy.* Accessed February 15, 2020. http://shardin.weebly.com/.

“Digital Content” definition1. Accessed February 15, 2020. https://en.wikipedia.org/wiki/Digital_content.

“Digital Content” definition2. Accessed February 15, 2020. https://magicmarketing.com.au/digital-content/.

“What is Reading?” *Teaching Reading.* Accessed February 15, 2020. https://www.tesol.org/docs/books/bk_ELTD_Reading_998.

“Reading” and “Digital Reading” definitions. Accessed February 15, 2020. https://www.merriam-webster.com/dictionary/read; https://dictionary.cambridge.org/dictionary/english/read; https://www.ukessays.com/essays/languages/definition-of-reading.phphttp://hipporeads.com/is-reading-dead-how-technology-and-literature-share-a-common-code/.

Mullan, Eileen. “What is Digital Content?” Accessed December 19, 2011. http://www.econtentmag.com/Articles/Resources/Defining-EContent/What-is-Digital-Content-79501.htm

About the Author:
Krisztina Szabó
https://orcid.org/0000-0001-8149-3684.

Foundations of the Social Futuring Index[1]

Zoltán Oszkár Szántó – Petra Aczél – János Csák – Chris Ball

Abstract

This paper presents a new, multidisciplinary concept called "Social Futuring" and introduces an index based on this concept, entitled the "Social Futuring Index". Settled into the intersection of philosophy, psychology, sociology, political theory and geopolitics among many other fields of social sciences social futuring and its application as an index addresses both academia and policymakers.

In the present article the concept is explained and then placed in the broader context of social sciences. We highlight that the most unique characteristic of social futuring is its fixed normative, analytical and discursive framework, the center of which is "a good life in a unity of order". Finally, we present the key elements of the index that are currently under construction.

Keywords: social futuring, social entities, Social Futuring Index, good life, normative standards.

1. INTRODUCTION

What is meant by "a good life in a unity of order" and what we expect a nation or country to provide for its citizens in terms of a good life is a question dating back at least to Ancient Greece. The traditional yet more modern approach simply looked at a country's GDP and assumed that GDP and welfare were closely related so that more GDP implied more human welfare. Today that approach is called into question from a range of intellectual perspectives, each generating its own branch of research around its specific area of critique. New measures have emerged to more completely capture the notion of "better", "welfare" and a "good life" from happiness indices to measures that incorporate environmental sustainability, all efforts to get a more complete picture.[2]

Each of those critiques brings a specific perspective, however. The happiness literature attempts to measure people's personal psychological wellbeing. Sustainability measurements focus on environmental wellbeing and long-term viability. Other indices focus on aspects of the political system like rule of law and others still continue to look at traditional economic indicators. But each function in isolation, in silos that are separate from each other, in an effort to better understand a particular aspect of society and social development.

[1] The present study is the updated and advanced version of the working paper entitled "The Concept and Measurement of Social Futuring" (Aczél et al. 2020). The authors express their gratitude to Pál Bóday, Eszter Deli, Judit Sebestény and Péter Szabadhegy for their valuable contribution to the final form of the paper.

[2] See Csák (2018) Introduction for greater detail about the concept of a "good life in a unity of order".

Social futuring represents a new, multidisciplinary approach that provides a holistic overview to measuring a social entity's ability to strategically plan for and sustain itself into the future while attempting to maintain the broad goal for its constituent members of achieving a "good life".

Environmental science is probably the furthest along in terms of obtaining widespread acceptance of the need to consider its modern critique on traditional measures of growth and wellbeing (Kocsis 2018). Sustainable economic development, for example, includes the environmental impact of economic development so that the environmental costs are incorporated into any economic cost-benefit analysis. The fundamental question being addressed by this is: how can we grow economically and yet also 'future proof' today's environment so that it is sustained – or even added to – for future generations.[3] From a process point of view, social futuring may be thought of as taking each discipline and asking how it can be made sustainable in the way that one future proofs a building or other physical object or system.[4]

Rather than treating each topic in a silo, however, social futuring attempts to bring their key insights under one roof and asks how this could be done for a society as a whole.[5] To do that, one first needs a common social goal against which to measure the current position and hence allow for a means to measure progress over time. As a first step, social futuring returns to the classical perspective of "a good life in a unity of order" as the broad notion of welfare in a society. It uses this as its normative metric and basis for evaluation and this normative framework is one of the aspects that makes social futuring a unique approach.

[4] For example, there is a great deal of literature on how we might measure happiness in societies (Helliwell, Layard and Sachs 2019). To apply the sustainability challenge here, one would ask something like the following: how can we 'future proof' a society's level of happiness so that its current level or even more happiness is sustainably maintained in order that future generations might too enjoy or improve upon it.

[5] Kocsis (2020) compared the Social Futuring Index with eight other country-level indices, namely with Better Life Index (BLI), Change Readiness Index (CRI), Global Resilience Index (GRI), Human Development Index (HDI), Happy Planet Index (HPI), Inclusive Development Index (IDI), Sustainable Development Goals Index (SDG), World Happiness Index (WHI) from three different aspects, such as Nature, Society, Economy. As a general result of this comparison he has concluded that SFI offers a balanced but fundamentally social composite for decision makers and those interested in the concept of futuring. Thus, both the concept of social futuring itself and the Social Futuring Index (SFI) based on it fill in the gaps in its economic-social-natural interest and complexity. All this may be even more evident if we consider the Aristotelian-Eudaimonic obligation evaluation of the index (Csák 2018) and an earlier version of its possible matrix-like, double grouping of its dimensions (Aczél et al. 2020, 35), which are not discussed here. Among the major composites known today, the SFI stands out primarily for its social (human) emphasis – while also taking into account economic-natural aspects in a proportionate way. This reflects the philosophy behind the indicator: the initial impulse of futureing is social, affecting the system of economic-natural relations. Calculating and tracking it can enrich future-oriented decision-making with new perspectives.
At the same time our complex approach has the special kind of limitation of not being centered around a specific sphere, but considering society as a whole, rooted in nature, while treating economy as embedded in society and culture.

After establishing the appropriate normative objective, social futuring must find its unique place in the approach of social sciences and then determine the means of measuring a social entity's progress toward its stated goal in reality. This is done through the Social Futuring Index (SFI)[6].

Social futuring is built on each of the key disciplines it incorporates. The Social Futuring Center (SFC) seeks to make field-specific research contributions around the concept of social futuring in the areas of philosophy, sociology, environmental and communication sciences, economics, future studies, geopolitics and political science. There is a need, however, to explain the core concept in a multidisciplinary way.[7]

This paper proceeds as follows. First, we present the key concept of social futuring. Second, we show that it is unique, and yet it incorporates elements of other well-established concepts. Finally, we present the key elements of the Social Futuring Index.

2. DEFINING THE CONCEPT OF SOCIAL FUTURING

The SFC defines social futuring as "a measure of a social entity's creative intent and potential to comprehend the ever-evolving world, its ability to get things done, to preserve and reproduce its way of life as well as to control its destiny in general" (Csák 2018, 22). This definition is broad enough to be applied to a wide range of social entities and yet precise enough to allow measurement. The definition starts with a "social entity", requires "intent" and a forward-looking approach along with an "ability" to make changes, all with a single goal in mind. To operationalize this concept, we next clarify each of these components.

2.1. SOCIAL ENTITY

The subject of social futuring is the social entity, "(…) an organism as understood based upon the concept of personhood, which denotes cognition, intentional activity and self-consciousness, as well as an awareness and recognition of the self's state of mind (as distinguished from others)" (Csák 2018, 24). Social futuring focuses on social entities constituted by persons who are given the ability to interpret things, make conscious decisions and take action and who are "embedded" into various groups and social networks. These include, but are not limited to, the following: organizations, settlements, regions, countries, country groups and potentially nations.

[6] The first SFI will be released in 2020 and will first focus on a country-level assessment. Subsequent efforts will then focus on ways to measure social futuring at more disaggregated levels, from cities all the way down to smaller organizations like companies, NGO's and associations.

[7] That is one of the main the purposes of the current paper, which was grounded by previous publications, describing the normative (Csák 2018), analytical (Szántó 2018) and discursive (Aczél 2018) framework of social futuring. While the previous publications considered these frameworks separately, the present one handles them in an integrated manner.

2.2. INTENT AND ABILITY

In order to qualify as a social entity capable of engaging social futuring, however, the social entity must meet five *necessary* conditions (NC). They are[8] that it

1. is able to operate functionally (NC1),
2. is able to sustain and reproduce itself over a long period of time (NC2),
3. is self-conscious (NC3),
4. is able to formulate an actionable strategy for itself (NC4), and
5. is able to provide its members with a "good life" (NC5).

The keys here are three: first, the entity must be able to manage itself over time. Second, it must be able to formulate a long-term goal for itself. NC1 and NC2 establish that an entity exists and functions over time. NC3 and NC4 establish that the entity is conscious and can establish its own goals. Finally, NC5 ensures that the entity can provide the "good life", which is, at a deeper level, the fundamental objective behind the whole notion of social futuring itself.

In many ways, the last condition, NC5, is also the starting point. If the entity is unable to provide its members with a "good life", either because it lacks resources or the requisite structure to plan and manipulate those resources (or for any other reason), then it will never be able to fully engage in social futuring in the sense we have in mind. The requirement that an entity be able to provide a "good life", in part or in entirety, restricts the types of entities we consider. For example, a city-planning group to build a bridge that is sustainable and future-proof would not count, but a city's mayor or planning group to manage the city over the coming years to improve the lives of its citizens would count.[9]

To understand the other conditions, we first turn to NC1 and NC2. A biological organism can meet NC1 and NC2. That organism can react to its environment over time, eat and store energy for the future, procreate etc. And, the broader forces of evolution will, through the entity's interaction with other entities and its environment, shape the organism today and shape it as a species over time. But we would not say that the organism ever engaged in social futuring because – to the best of our knowledge – it never became self-aware in a personhood and a social sense and it never defined its own long-term goals upon which it then acted. That is, the organism and its species lacked NC3 and NC4. Likewise, if a few people decide to form a club, they may pick a name for the club, define its membership and even establish its goals. These would meet NC3 and NC4, but until the club becomes a viable entity that can actually manipulate resources to maintain itself over time (i.e., meets NC1 and NC2), we cannot say that the club engaged in or can engage in social futuring. So, the entity must be "social" and self-aware. It must also be able to make a strategic plan for itself and be able to carry it out to some extent.

[8] Note that this list is a modified version of the one found in Szántó (2018).

[9] We leave the topic of what exactly the "good life" is for section 2.4. below, since the concept is deeply connected with the normative framework of social futuring.

2.3. FORWARD LOOKING

The ability to imagine the future, to progress towards the future and to arrange future possibilities are distinctive features of humans. This ties in both with the definition of social futuring as dealing with the future and with social entities being constituted by people who are distinct biological forms defined historically and philosophically on the basis of the notion of personhood. Furthermore, it is quite logical that if a group of people are to set long-term objectives for themselves, they must be forward looking. This is therefore one of the more obvious and logical necessary requirements for an entity to be able to engage in social futuring, essentially NC2 and NC4 in the above list.

2.4. THE NORMATIVE GOAL AND FRAMEWORK

All forms of welfare analysis must assume *a priori* a normative measure against which one can measure improvement or lack thereof. Economists assume people maximize utility, which is an individual-specific ranking of alternative outcomes. If utility is higher, then economists claim welfare has improved. But it has long been recognized and formally shown by Kenneth Arrow (1950), that aggregating utility is notoriously difficult if not entirely impossible in practice. As a result, many in the social sciences seek alternative measures of aggregate or proxies for wellbeing such as happiness, freedom, GDP frequently, equality and so on. In the end, if we want to measure progress, we need to assume the goal toward which progress is made.

The social futuring initiative assumes a broad definition that is grounded in the moral philosophical Aristotelian-Thomist tradition, which considers that "we are in some respects social beings, a genuine aspect of whose telos is participation in shared ends" (Haldane 2009, 231-232). The social futuring project is about the study of characteristics that make this *telos* more or less successful and starts with the assumption that the ultimate purpose of social entities is to enable a good life that is worth preserving and reproducing. Therefore, maintaining the "good life in a unity of order" is the starting place and ultimate normative objective for social futuring.

The notion of "the good life" is broad in the way that "utility" is broad for economists. Different societies and social entities may define the good life differently for themselves. As a matter of fact, NC3 and NC4 require that the social entity be able to define the good life for itself. Therefore, there is not a single definition like more happiness or GDP or consumption that the social futuring project or index relies on to measure "good".[10] The "unity of order" provides the requirement that the persons in the social entity are indeed part of the social entity itself. This returns us to NC3 and NC4 which together argue that the individuals that collectively constitute the social entity are self-conscious as a group and themselves constitute the group. Based on these insights, in order to opera-

[10] This allows the SFI eventually to consider the cases of smaller entities like a company, association or church that might define good and wellbeing for its members very differently from another company, association or church. Likewise, cities might define "good" differently than countries and different countries might define it differently from each other.

tionalize the normative framework, the SFC established the following normative standards:[11]

- *Peace and security*: This is the minimum substance of a „unity of order". It enables social entities to reproduce, to raise children and to provide for themselves and others in a safe environment, furthermore to make predictions, to set goals and functionally influence their future operation based on strategic assets.
- *Attachment:* This is essential for healthy bodily, psychological, intellectual and spiritual human development. The most basic unit of attachment is the family, which determines the consciousness of what a "relationship, dignity, equity, authority and hierarchy are; what is good and bad, just and unjust; what is love, gift and reciprocity" (Csák 2018, 37), however, patriotism and spirituality are also key dimensions of the standard.
- *Care* (material advancement and freedom): "The maintenance of material goods must entail the accepted practices of production, distribution and acquisition; use and disposition of private or public goods; extendable management skills; and, therefore an image of wealth and the nature of work" (Csák 2018, 37-38). Freedom is the ability of self-determination and self-reliance to actualize one's potential and capacity to control their fate.
- *Balance:* This is a state of mind, an attitude towards life that reflects the equilibrium between the concern for the self and the care about others – that is, next generations. It is thus a prerequisite of the compound of wellbeing and generativity. Balance is about being free of unproductive societal comparisons and having the balance to give, lead and fulfil human life.

These four normative standards follow each other in a hierarchical order, meaning that without the minimum level of peace and security no attachment, care and balance is possible. Without the minimum level of attachment, no care and balance is possible. And last but not least, without care balance is also impossible.

2.5. MUST ALL CONDITIONS BE MET?

Sufficient Conditions and Partial Results

Of course, meeting all necessary conditions, 1-5, defines the ideal and complete Social Futuring entity. In this sense NC1-NC5 are sometimes referred to as conjunctive prerequisites in that all five must be met simultaneously for an entity to be considered fully to engage in social futuring. But there are different levels, degrees or forms of social futuring that we might also consider when entities engage in some degree of ensuring their own future viability.

The disjunctive *sufficient* condition for the future viability of any social entity are that it be able[12]

- to bring about changes, and to prepare for influencing expected changes,
- to prepare to exploit the opportunities and neutralize the limitations of the expected changes and,

[11] See Csák (2018) for greater detail.

[12] See Szántó (2018) for greater detail on these conditions and their implications for social entities.

- to prepare to manage the risks associated with the expected changes.

The implication of these looser, disjunctive conditions is that there can exist various forms or levels of social futuring in which an entity can engage, while still being considered as social futuring and not just planning. The result is that there are three broad categories of social futuring:

- *Proactive* occurs when social entities seek to understand, bring about and influence the changes that are expected in the future. This is the most complete form and closest to complete social futuring.
- *Active* occurs when the possible agents of social entities are prepared to counteract the limitations and/or to take advantage of favorable opportunities of future change.
- *Reactive* occurs when social entities strive to manage the risks that accompany change.

3. PLACING THE CONCEPT IN BROADER CONTEXT

3.1. TRADITIONAL SOCIAL SCIENCES

The distinction is most clear by starting with the social science most distant from social futuring. That science is economics. Economics, since at least the time Adam Smith's "invisible hand"[13] was formalized, studies almost the exact opposite of what social futuring aims to study. Social futuring examines the success of self-aware collective groups called social entities that define and strategically move toward their collective goal. Economics studies how self-interested individuals manage to organize limited resources without a central design through a spontaneous ordering subject only to the natural laws of economics. In the words of Friedrich Hayek[14] "…economics has come nearer than any other social science to … show that … the spontaneous actions of individuals will, under conditions which we can define, bring about a distribution of resources which can be understood as if it were made according to a single plan, although nobody has planned it, seems to me indeed an answer to the problem which has sometimes been metaphorically described as that of the "social mind" (Hayek 1937, 52). And elsewhere, more succinctly, he states "[t]he economic problem of society is … a problem of the utilization of knowledge which is not given to anyone in its totality" (Hayek 1945, 520).

[13] Adam Smith ([1776] 1977, 421): "By directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it."

[14] The full quote is "…economics has come nearer than any other social science to an answer to that central question of all social sciences: How can the combination of fragments of knowledge existing in different minds bring about results which, if they were to be brought about deliberately, would require a knowledge on the part of the directing mind which no single person can possess? To show that in this sense the spontaneous actions of individuals will, under conditions which we can define, bring about a distribution of resources which can be understood as if it were made according to a single plan, although nobody has planned it, seems to me indeed an answer to the problem which has sometimes been metaphorically described as that of the "social mind"." (Hayek 1937, 52).

Economics starts by considering a single individual or a collection of individuals, each of whom form their own private and separate plans. They do not have a common plan and the economic question then becomes an exploration how these individuals manage to achieve so much without a common plan. Mancur Olson (1965) goes so far as to argue in his foundational book, *The Logic of Collective Action: Public Goods and the Theory of Groups*, that studying "collective action" requires understanding that even if self-interested individuals agree on a common interest, the group they form will not represent those interests by acting in some group-interest (Olson 1982, 17). He argues that "large groups, at least if they are composed of rational individuals, will not act in their group interest" (Olson 1982, 18).

Thus, a Hayekean-conceived economic order, or social entity, cannot engage in social futuring any more than the biological organisms mentioned earlier can. Such entities fail on necessary conditions NC1 and NC4 for sure and possibly NC2 as well, depending on how we define it.

The economic approach subsequently influenced political science as well, infusing it with an individualistic, Hayekean foundation. "The importance of Olson's argument to the history of social science cannot be overestimated. Prior to Olson, social scientists typically assumed that people would instinctively or naturally act on common interests, and that inaction needed to be explained" (Oliver 1993, 273). "After Olson, most social scientists treat collective action as problematic. That is, they assume that collective inaction is natural even in the face of common interests, and that it is collective action that needs to be explained" (Oliver 1993, 273-274).

A range of modern social scientists, even in relatively traditional fields, have however begun to adopt alternative approaches. Easily included in this list could be Harari's recent contributions to rethinking both human history and human future as in his works *Homo Deus: A Brief History of Tomorrow* (Harari 2017) and *21 Lessons for the 21st Century* (Harari 2018), where he merges a long-term, macro-historical view with insights into human evolution to address the concerns all humans are facing and will face in the future. A similar, forward-looking approach, applied a little less broadly than in Harari's exceptionally wide brush strokes, would be the work of George Friedman generally focusing on global geopolitical trends, best captured in print in *The Next 100 Years* (Friedman 2009). A final approach, applied to a cross section of human behavior, but not necessarily across time or with an eye toward the future, would be *Bursts* by Albert-László Barabási (2010).

The conclusion here is that – despite some recent innovations from those working in the vein of Barabási, Friedman and Harari – most traditional social sciences follow the economic approach of considering individual rational actors pursuing their own self-interest. The starting point is to consider individuals who have their own, not common plans. Social futuring, by way of contrast, starts by only considering a collection of individuals who have a common plan and then studies how that collective group achieves a broader outcome as defined by their plan.

3.2. NEW SOCIAL SCIENTIFIC APPROACHES

There are other branches of the sciences that have gained prominence as separate fields in recent years. These fields share much more in common with social futuring and reveal that the intellectual location of social futuring is more in line with these newer approaches.

They are the study of resilience, future orientation and future proofing. Comparing them with social futuring helps clarify the areas social futuring shares with, or builds upon them and where it is distinct from them which is also summarized in Figure 1.

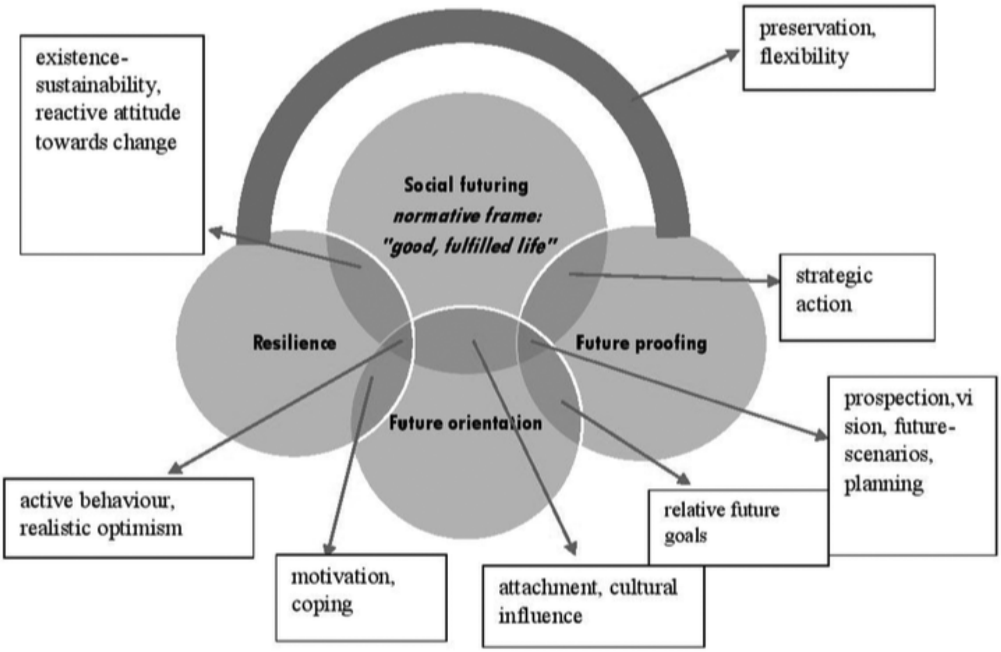

existence-sustainability, reactive attitude towards change
preservation, flexibility
Social futuring
normative frame: "good, fulfilled life"
strategic action
Resilience
Future proofing
Future orientation
prospection, vision, future-scenarios, planning
active behaviour, realistic optimism
relative future goals
motivation, coping
attachment, cultural influence

Figure 1: Overlapping and distinct elements of social futuring. From Aczél (2018, 71).

3.2.1. Resilience

Disciplines like physics, ecology and psychological discourse use the term resilience to mean flexible, beneficial adaptation to traumas, stress and difficulties, which occasionally involves the process of learning and development.[15] The first and perhaps biggest distinction between the concept (and study) of resilience and that of social futuring is that resilience lacks a normative framework other than the objective of "allowing something to persist". A secondary distinction is that resilience generally views change as a negative influence to be resisted, while change is an opportunity for social entities engaged in social futuring, since it is necessary for them to achieve their long-term objectives.

To some extent, social futuring also includes the concept of resilience to the extent that it includes as a central issue preserving, protecting and reproducing "the good life"

[15] Aczél (2018, 54) reviewed some "(…) tests and indexes that have been developed to measure personal and age-related resilience (the Connor–Davidson Resilience Scale, the Response to Stressful Experiences Scale, the Dispositional Resilience Scale-15, the Resiliency Scale for Children and Adolescents, RSCA Global Scales and Index) use self-reporting or assessments primarily to find out how people cope with the challenges of reactivity, assertiveness, attachment, control and problems, each of them considered a factor in resilience."

for its constituent members. In this sense, social futuring entities must identify a core identity that is made resilient while planning long-term for broader changes in an adaptive, evolutionary sense.[16]

3.2.2. Future Orientation

Future orientation intends to capture the degree to which an individual thinks in advance as well as capture his/her attitude regarding the future and how it connects to the present and past (Aczél 2018, Monda 2018). Cultures may differ on their perspective of time, whether it is linear or not and the degree to which it may be manipulated. Disciplines also differ in their perspective on time. People in more technologically-oriented disciplines and societies, for example, are more focused on performance, completion and achievement over time so that the future becomes measured in terms of performance generally.[17]

Based on Trommsdorff (1983), the concept of future orientation can be interpreted as an attitude of humans (and culture) referring to the future. It "expresses the mindset through which the conception of the future appears, and lastly it is used to mean such culturally and individually determined complex behaviors which contribute both to culture and to the individual and in which we can suppose a future orientation" (Aczél 2018, 64). Social futuring inherently includes future orientation, since it is primarily about the future itself. While it is certainly necessary for a social entity that engages in social futuring to have a future orientation, social futuring itself is about strategic action extending forward in time while future orientation is simply a matter of whether or not the entity looks forward and, if so, how far into the future[18].

3.2.3. Future Proofing

Future proofing is a concept that has become much more common in technological and architectural industries. The core concept is that an investment into a product, be it a smart phone or a building, only makes sense to the extent that the generated product is sufficiently future proofed to survive long enough to provide a sufficient return on investment. In the case of a technology-based product, the threat comes from competitors developing new technologies that make current products/technologies obsolete. In the case of architecture, there is a technological component, but more importantly, the physical structure needs to withstand environmental forces for a meaningful period of time.

[16] For this reason, Figure 1 shows the intersection of the two concepts as representing the common elements of "existence-sustainability and a reactive attitude towards change".

[17] Aczél (2018, 65) summarized The Future Orientation Index in the following way: it "explores future orientation using trends in information seeking by looking at Google searches for specific years written in Arabic numbers. The FOI expresses the extent to which internet users worldwide (by country) in a given year are more interested in information available from upcoming than previous years."

[18] As shown in Figure 1, the two do share the fact that people's attitudes and understanding of the future are heavily influenced by their culture as well as their attachment to the present and their core beliefs. As in the case of resilience, the biggest difference again is that social futuring starts from the premise of a defined social entity with a set normative framework and objective, whereas future orientation is entity-less and essentially non-normative in nature.

Therefore, we conclude that the essence of "(...) future proofing is that investors should prevent the creation of new technologies that are unfit for improvement and they should rather promote the creation of flexible open-ended systems that adapt to changing needs" (Aczél 2018, 69). The concept of future proofing, then, refers to the logic of informed strategic formulation and development that rest on well-grounded foresight. In the case of organizations, however, future proofing can be considered a given future-oriented way of promoting common thinking. Social futuring is, at one level, most similar to the concept of future proofing (as compared to resilience or future orientation). One can almost think of social futuring as the future proofing of a given social entity's values and goals for its constituent members. As a result, they have in common that both are concerned with strategic action, have a vision for the future and, combining these two, necessitate some degree of planning.

The two concepts differ radically, however, in their normative basis and on their areas of focus. Firstly, future proofing has no normative basis other than survival of the current state for as long as possible whereas social futuring starts be establishing a normative framework and goal, that of "maintaining the good life in a unity of order for its constituent members". Secondly, future proofing tends to be an industry-specific concept. That is, it has a very different meaning for each specific technological industry, since their competitors are different, while social futuring aims precisely to develop a common framework of analysis that can be used consistently across individual social entities, including businesses. Moreover, the concept of social futuring can also be much broader by considering very large social entities such as countries[19].

4. THE SOCIAL FUTURING INDEX (SFI)

The study of resilience, future orientation and future proofing contribute new insights into how cultures differ and what parameters affect an individual's or a group's ability to engage the world around them over time. Social futuring aims to do the same while providing a normative framework for analysis. But, as a project, it is not merely an intellectual endeavor. The social futuring initiative set the practical goal of developing the SFI, a composite measure of countries comprising a number of dimensions and indicators in four pillars. The indicators of the index are selected from a number of internationally recognized databases which are provided by OECD, World Bank, World Value Survey etc. The focus of the Index is a 'life in a unity of order', which can be characterized by the aforementioned four normative standards, namely peace and security, attachment, care (material advancement and freedom) and balance, as it is visualized in Figure 2.

[19] The summary of the comparison and contrast of social futuring versus these other views can be found in both Figure 1 and in Table 1 (in the Appendix). Table 1 presents a more nuanced view of the differences breaking each concept into the components of its views on disruption, risk, process, view on opportunities, whether it is primarily reactive, active or pro-active, whether it is primarily focused on the individual or society, and whether it is motivated to change via incentives or more strategic in nature. Her conclusion is that social futuring includes all the categories of the other concepts except one: disruption. Otherwise, in many regards, social futuring is the larger category or umbrella, building on the other concepts.

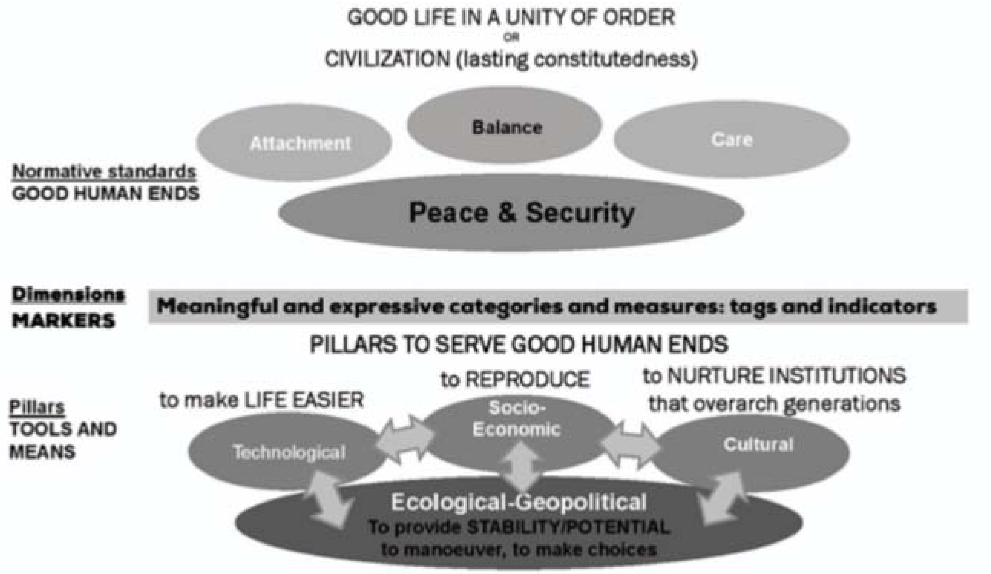

GOOD LIFE IN A UNITY OF ORDER
OR
CIVILIZATION (lasting constitutedness)
Attachment
Balance
Care
Normative standards
GOOD HUMAN ENDS
Peace & Security
Dimensions
MARKERS
Meaningful and expressive categories and measures: tags and indicators
PILLARS TO SERVE GOOD HUMAN ENDS
to REPRODUCE
to NURTURE INSTITUTIONS that overarch generations
Pillars
TOOLS AND MEANS
to make LIFE EASIER
Socio-Economic
Technological
Cultural
Ecological-Geopolitical
To provide STABILITY/POTENTIAL to manoeuver, to make choices

Figure 2: The conceptual interrelations of the SFI's normative standards, dimensions, and pillars

The scores of the Index will be interpreted from the perspective of the worthwhile life as a standard.

The notion that an approach should be measurable and should provide a benchmark for progress, is not unique within the field of social sciences. Indeed, traditional social sciences have developed growth indices and institutional indices important to growth, freedom and the rule of law.[20] The newer areas of study like that of resilience, future orientation and future proofing also developed indices in their specific fields.[21]

While the ultimate aim is to develop generally applicable indices for social entities of all types and sizes, the social futuring project started by first focusing on developing a country-level index for three practical reasons. First, a country is about the largest social entity that has a defined leader (the government or state) that represents the constituent members, generally through democratic institutions. Second, there are existing data on multiple countries, allowing the first indices to be constructed from current data sources

[20] As examples, see the World Bank Development Indicators (World Bank 2019), or the Heritage Foundation Freedom Index (Heritage 2019), or the CATO Human Freedom Index (Vásquez and Porcnik 2018).

[21] For resilience, either of individuals or larger aggregates of individuals, there are: the Connor–Davidson Resilience Scale, the Response to Stressful Experiences Scale, the Dispositional Resilience Scale-15, the Resiliency Scale for Children and Adolescents, RSCA Global Scales and Index (Prince-Embury 2008, Prince-Embury and Saklofske 2012). For future orientation there is now The Future Orientation Index (Preis et al. 2012). Since future proofing is an industry specific matter, there are myriad industry specific metrics employed that conform to each industry's regulatory standards or are proprietarily developed to respond to competition.

rather than requiring that the research project solve two problems at once: constructing an index as well as generating new data. Third, in the same way that the concept of social futuring needed to define itself in comparison to other concepts or approaches in the social sciences, so too must an index find its home in contrast to other existing indices. Therefore, starting with countries that are part of other currently existing indices allows the SFI to distinguish itself by highlighting the differences and similarities to other, regularly published indices.[22]

The outlines of the SFI are presented in Figure 3 and summarized here, in order to further conceptualize the SFI and the pillars of the Index implemented by the SFC. According to this logic, the concept for the index is based around the following four pillars:

- Ecological-Geopolitical,
- Technological,
- Socio-Economic, and
- Cultural.

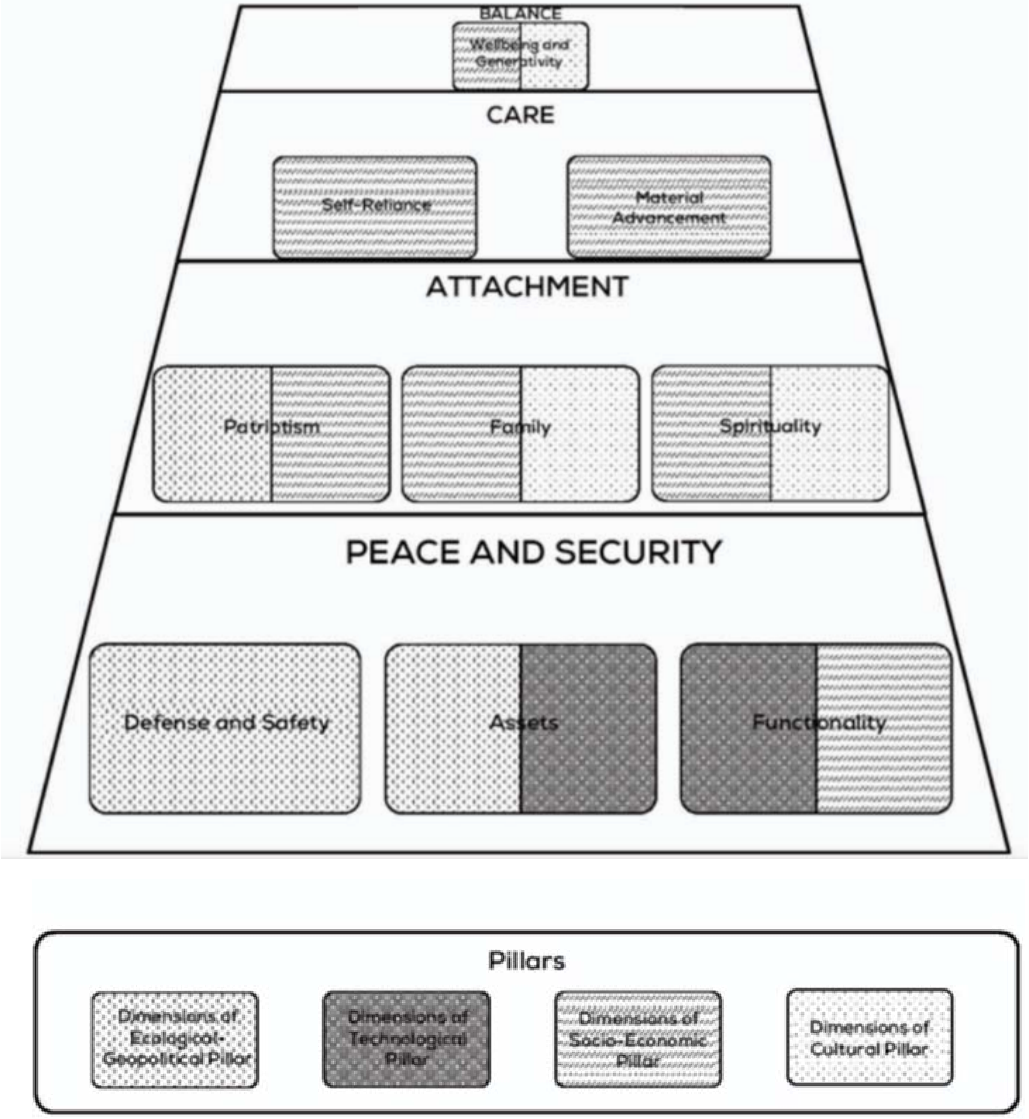
The *Ecological-Geopolitical* pillar captures aspects of a social entity such as its basic assets (energy, water, land etc.) without which it would not have resources to maintain itself. Moreover, it includes dimensions such as measures of patriotism, defense and safety to capture various aspects of belonging to the social entity as well as the assets/resources needed to engage in social futuring. The *Technological* pillar includes aspects such as a social entity's ability to network/connect, innovate and function generally. Basic functioning requires fundamental resources like clean water, while innovation includes a need for a legal framework for patents and intellectual property. Finally, the ability to network and connect can be measured physically, such as roads or digitally, such as internet access, ICT use. The *Socio-Economic* pillar includes classical economic areas like capital, labor and various expenditures as well as indicators of unemployment, schooling and GDP/capita. Socially, the core unit considered for a stable socially cohesive society that engages in social futuring is the family and therefore the SFI includes measures such as fertility, the number of single-parent households, couples with children, work-life balance, ageing and inequality. Finally, the *Cultural* pillar – in many ways the single dimension that makes the SFI unique, since its normative basis is one of the key aspects making the concept of social futuring itself unique – includes measures such as religiosity and following traditions.

As a result the four pillars and four normative standards outline nine dimensions:[23]

[22] This last reason also allows us to test statistically for the difference between the SFI and other indices, adding an objective element to the claim that the SFI is unique.

[23] See Table 2 (in the Appendix) for the definitions and conceptualization of each dimension.

Figure 3: The normative standard based matrix structure of the SFI
SOCIAL FUTURING INDEX
Good Life in a Unity of Order


BALANCE
Wellbeing and Generativity
CARE
Self-Reliance
Material Advancement
ATTACHMENT
Patriotism
Family
Spirituality
PEACE AND SECURITY
Defense and Safety
Assets
Functionality
Pillars
Dimensions of Ecological-Geopolitical Pillar
Dimensions of Technological Pillar
Dimensions of Socio-Economic Pillar
Dimensions of Cultural Pillar

Within each pillar and dimension of each normative level, the SFI includes multiple indicators. Each is weighed/ranked to provide sub-indices and then aggregated to form the overall ranking. This allows one to disaggregate the overall ranking to see where any specific country is relatively stronger or weaker. It provides information and potential guidance for countries wishing to improve their own social futuring efforts.

5. CONCLUSION

This paper has presented the holistic concept of social futuring and the foundations of the Social Futuring Index. We first explained the basis for the definition of social futuring and argued that it is a conceptually unique approach in social sciences. We then showed where it fits within modern approaches to thinking about societies and the future. The element that was most consistently found to make the concept unique is that it is founded within a specific normative framework. The second most important element, especially separating it from traditional social sciences, was that the starting point of analysis is the social group or entity, which presupposes self-conscious and self-constituting social entities that share a common purpose. Finally, we elaborated on the general framework of the index, based on four normative standards, four pillars, and the nine dimensions they co-create.

According to our intentions, the concept of Social Futuring and the SFI may be of interest for the Academia, especially for those economists and social scientists who are sensitive towards the holistic, multi-disciplinarian and complex approaches of thinking about good life today and tomorrow. However, policy- and decision-makers may also benefit from the findings of the SFI. During the interpretation and dissemination of our country-level results, we will also focus on the practical applicability of our index, providing the distinction of the so called policy-sensitive indicators among the indicators our index takes into account. According to the information stemming from them, our best hope is that the points of intervention could easily be identified in different policy areas as well.

BIBLIOGRAPHY

Aczél, Petra. "Social Futuring – A Discursive Framework." *Society and Economy* 40, Issue S1 (2018): 47-75. https://doi.org/10.1556/204.2018.40.s1.4.

Aczél, Petra, Chris Ball, János Csák, and Zoltán O. Szántó. "The Concept and Measurement of Social Futuring." Working Paper Series no. 8 (2020): CUB, SFC.

Arrow, Kenneth Joseph. "A Difficulty in the Concept of Social Welfare." *The Journal of Political Economy* 58, no. 4 (1950): 328-346. https://doi.org/10.1086/256963

Barabási, Albert-László. *Bursts: The Hidden Patterns Behind Everything We Do*. New York, NY: Penguin Group, 2010.

Csák, János. "Social Futuring – A Normative Framework." *Society and Economy* 40, Issue S1 (2018): 21-45. https://doi.org/10.1556/204.2018.40.s1.3

Friedman, George. *The Next 100 Years: A Forecast for the 21st Century*. New York, NY: Random House, 2009.

Haldane, John. *Practical Philosophy: Ethics, Society and Culture*. St. Andrews Studies in Philosophy and Public Affairs, Imprint Academic, 2009.

Harari, Noah Yuval. *Homo Deus: A Brief History of Tomorrow*. New York, NY: HarperCollins, 2017.

Harari, Noah Yuval. *21 Lessons for the 21st Century*. New York, NY: Random House, 2018.

Hayek, Friedrich August von. "Economics and Knowledge." *Economica* New Series 4, no. 13 (1937): 33-54.

Hayek, Friedrich August von. "The Use of Knowledge in Society." *The American Economic Review* 35, no. 4 (1945): 519-530.

Heritage. *2019 Index of Economic Freedom*. The Heritage Foundation, 2019. Accessed July 7, 2019. https://www.heritage.org/index/

Helliwell, John F., Richard Layard, and Jeffrey D. Sachs. *World Happiness Report 2019.* United Nations, 2019.

Kocsis, Tamás. "Finite Earth, Infinite Ambitions: Social Futuring and Sustainability as Seen by a Social Scientist." *Society and Economy* 40, Issue S1 (2018): 111-142. https://doi.org/10.1556/204.2018.40.S1.6

Kocsis, Tamás. "The Social Futuring Index (SFI) in the Context of Economy, Society and Nature: Intenscoping Nine Composite Indices Measuring Country Performance." 2020. (Unpublished manuscript)

Monda, Eszter. "Social Futuring in the Context of Futures Studies." *Society and Economy* 40, Issue S1 (2018): 77-109. https://doi.org/10.1556/204.2018.40.S1.5

Oliver, Pamela E. "Formal Models of Collective Action." *Annual Review of Sociology* 19, (1993): 271-300.

Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Boston, Mass: Harvard University Press, 1965.

Olson, Mancur. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidites.* New Haven, CT: Yale University Press, 1982.

Preis, Tobias, H. S. Moat, H. E. Stanley, and Steven R. Bishop. "Quantifying the Advantage of Looking Forward", *Nature/Scientific Reports* 2 (2012). https://doi.org/10.1038/srep00350

Prince-Embury, Sandra. "The Resiliency Scales for Children and Adolescents, Psychological Symptoms, and Clinical Status in Adolescents." *Canadian Journal of School Psychology* 23, Issue 1 (2008): 41-56. https://doi.org/10.1177/0829573508316592

Prince-Embury, Sandra, and Donald H. Saklofske (eds.). *Resilience in Children, Adolescents and Adults: Translating Research into Practice.* New York, NY: Springer, 2012.

Smith, Adam. *The Wealth of Nations.* Chicago, IL: The University of Chicago Press 1977 [1776], (edited by Edwin Cannan).

Szántó, Zoltán O. "Social Futuring – An Analytical Conceptual Framework." *Society and Economy* 40, Issue S1 (2018): 5-20. https://doi.org/10.1556/204.2018.40.S1.2

Trommsdorff, Gisela. "Future Orientation and Socialisation." *International Journal of Psychology* 18 (1983): 381-406. https://doi.org/10.1080/00207598308247489

Vásquez, Ian, and Tanja Porcnik. *The Human Freedom Index 2018: A Global Measurement of Personal, Civil, and Economic Freedom.* The Cato Institute, the Fraser Institute, and the Friedrich Naumann Foundation for Freedom, 2018.

World Bank. *World Development Indicators.* The World Bank Group, 2019. Accessed July 7, 2019. http://datatopics.worldbank.org/world-development-indicators/

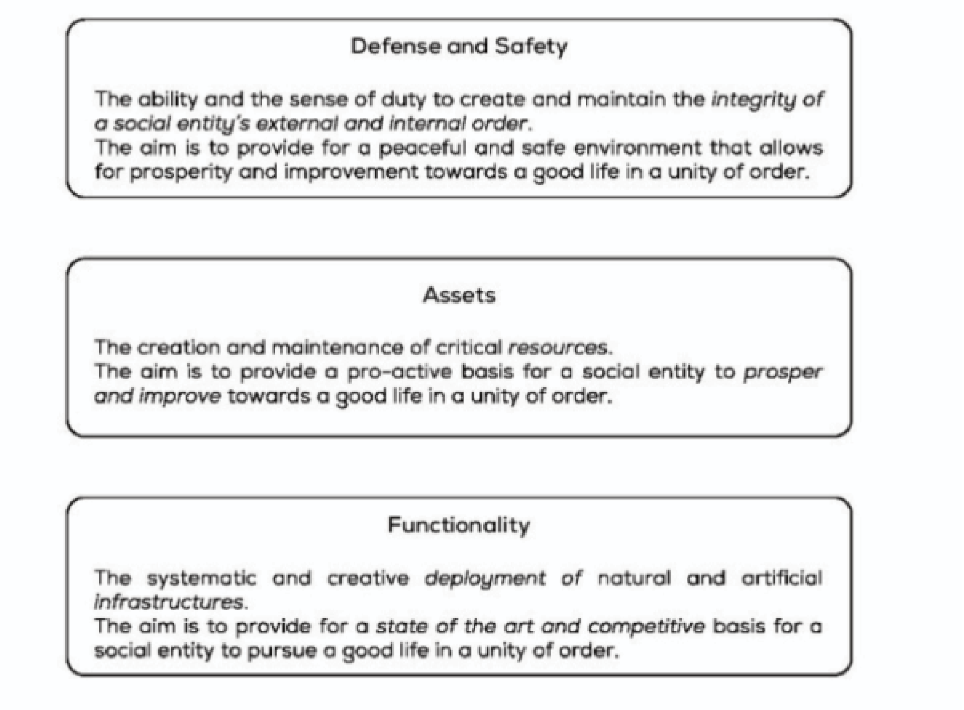
Author information
Zoltán Oszkár Szántó, Corvinus University of Budapest (CUB)
https://hu.linkedin.com/in/zoltán-oszkár-szántó-53b15510
Petra Aczél, Corvinus University of Budapest (CUB)
https://hu.linkedin.com/in/petra-aczél-2126b28b
János Csák, Corvinus University of Budapest (CUB)
Chris Ball, Quinnipiac University, Invited Research Fellow at CUB,
https://www.linkedin.com/in/christopher-ball-b191b051

APPENDIX

Table 1:
Comparison of Social Futuring, Resilience, Future Orientation and Future Proofing. From Aczél (2018).

	Conception of change			Attitude to change			Vision as a condition	Entity/agency				Action	
	Disruption, risk	Process	Opportunity	Reactive	Active	Pro-active		Individual	Social	Cultural	Instrumental	Motivated	Strategic
Social futuring		X	X	X	X	X	X		X	X	X	X	X
Resilience	X			X	X			X	X			X	
Future orientation		X			X	X	X	X		X		X	X
Future proofing	X		X		X	X	X				X		X

Table 2: The definitions of the dimensions of the SFI



Defense and Safety

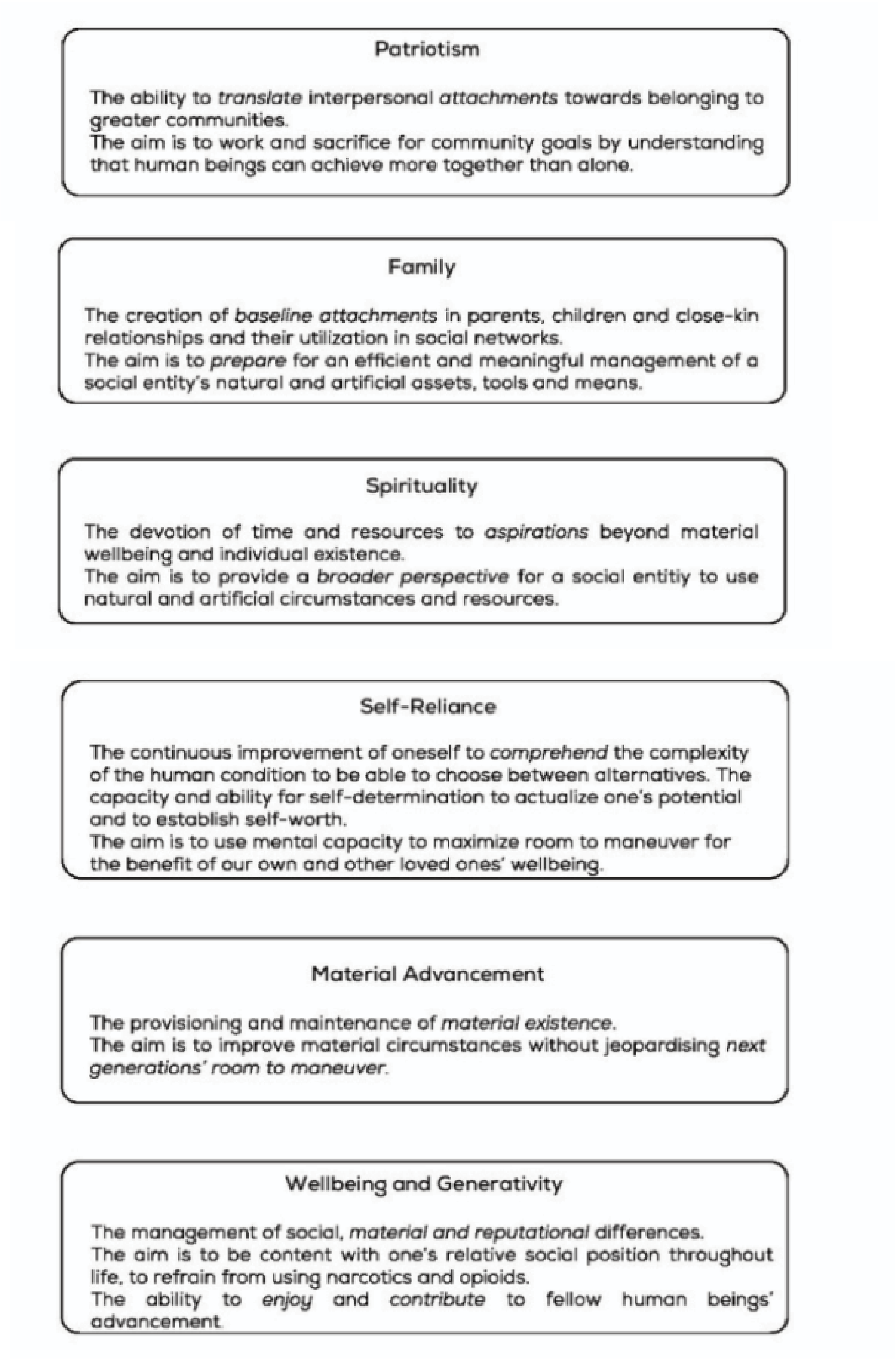
The ability and the sense of duty to create and maintain the *integrity of a social entity's external and internal order*.
The aim is to provide for a peaceful and safe environment that allows for prosperity and improvement towards a good life in a unity of order.

Assets

The creation and maintenance of critical *resources*.
The aim is to provide a pro-active basis for a social entity to *prosper and improve* towards a good life in a unity of order.

Functionality

The systematic and creative *deployment of* natural and artificial *infrastructures*.
The aim is to provide for a *state of the art and competitive* basis for a social entity to pursue a good life in a unity of order.



Patriotism

The ability to *translate* interpersonal *attachments* towards belonging to greater communities.
The aim is to work and sacrifice for community goals by understanding that human beings can achieve more together than alone.

Family

The creation of *baseline attachments* in parents, children and close-kin relationships and their utilization in social networks.
The aim is to *prepare* for an efficient and meaningful management of a social entity's natural and artificial assets, tools and means.

Spirituality

The devotion of time and resources to *aspirations* beyond material wellbeing and individual existence.
The aim is to provide a *broader perspective* for a social entitiy to use natural and artificial circumstances and resources.

Self-Reliance

The continuous improvement of oneself to *comprehend* the complexity of the human condition to be able to choose between alternatives. The capacity and ability for self-determination to actualize one's potential and to establish self-worth.
The aim is to use mental capacity to maximize room to maneuver for the benefit of our own and other loved ones' wellbeing.

Material Advancement

The provisioning and maintenance of *material existence.*
The aim is to improve material circumstances without jeopardising *next generations' room to maneuver.*

Wellbeing and Generativity

The management of social, *material and reputational* differences.
The aim is to be content with one's relative social position throughout life, to refrain from using narcotics and opioids.
The ability to *enjoy* and *contribute* to fellow human beings' advancement.

A Theory of (Sexual) Justice: the roboethician's edition

Radu Uszkai

Abstract

Sex robots have been gaining significant traction in the media and in pop culture. Each new launch of an updated model or a new entrepreneurial innovation on the sex robot market was signaled and discussed at length in the media. Simultaneously, Hollywood productions and popular TV series have graphically illustrated and brought forth serious questions regarding human – sex robot relationship. Unsurprisingly, philosophical interest is already extensive, with a series of papers and books tackling a wide array of issues related to sexbots. The purpose of my paper is that of exploring one potential deployment of sex robots: as a solution for addressing claims of sexual justice. I will begin with a short overview of the debate regarding sex rights for people with disabilities and argue that a Rawlsian account of sexual justice is possible. One of the main claims of the paper will be that there might be a strong link between sex rights and Rawlsian primary goods. I will then argue that, from a Rawlsian framework, it makes sense to adopt an anthropocentric meta-ethical approach to human – sex robot interactions. In the last part of the paper, I will present and criticize the main objections that have been brought against the manufacture and selling of sex robots. Even assuming that the objections were correct, they do not hold in the case of the use of sex robots by people with mental or physical disabilities.

Keywords: roboethics; sex robots; Rawls; free market fairness; sexual justice; sex rights

1. On the idea of sex rights

During the past decade the *Journal of Medical Ethics* was the host of a debate on the idea of sex rights for the disabled. The spark was a paper written by Appel (2010) in which he argued that we have focused almost exclusively on protecting vulnerable groups from abuse and largely ignored the intimacy needs of people with either physical or mental disabilities. His contention is that people have both positive and negative sex rights and that they "encompass the right to experience pleasurable sexuality, which is essential in and of itself and, at the same time, is a fundamental vehicle of communication and love between people" (152). The distinction between negative and positive rights goes back to Isaiah Berlin's (2002) distinction between negative and positive liberty. Negative rights carve out areas in which we are free from any type of coercion from the state or interference from society so long as we ourselves do not interfere with the negative rights of other individuals. Thus, having a negative right to X means that no one should interfere with my having access to X, acquiring X or enjoying X. On the other hand, positive rights, just like Berlin's positive freedom, are rights to be provided with X if X increases your autonomy and you are unable (due to a wide variety of objective reasons) to have access to X on your own.

The implications of such a normative position are twofold. Taking negative sex rights seriously would entail that we should reform policies in nursing facilities so as to allow sexual intercourse on their premises. A major and highly contentious one (for some moral philosophers) has to do with prostitution. If it is true that people have negative sex rights, then we should at least carve out exceptions for buying sex and legalize (within a specific scope) prostitution. There is also a case to be made in favor of publicly subsidizing prostitution for the disabled if they lack the resources to buy such services, as sex rights are also positive rights.

In contrast to Appel, Di Nucci (2011) has a largely skeptical and critical approach. If people would have such a positive right to sexual satisfaction, then this would deprive others of their negative rights: "universal positive sexual rights are incompatible with universal negative sexual rights. If A has a positive sexual right, then that means that there is at least one person who would lack negative sexual rights. Namely the person who would be supposed to fulfill A's positive sexual rights. If everybody has negative sexual rights, then everybody has the right to refuse to fulfill A's sexual needs, but then A has no positive right to sexual pleasure." (159). A better solution for providing sexual satisfaction for people with disabilities would circumvent the ethical issues associated with both public subsidies and legalizing (albeit partially) prostitution. Such a solution could be found, as strange as it may sound, in establishing "charitable non-profit organizations, whose members would voluntarily and freely provide sexual pleasure to the severely disabled" (2011, 160).

A more promising approach to the issue resides in Thomsen (2015). Sex rights are especially important for people who are, in his words, "relevantly disabled": persons who have sexual needs and desires that are difficult to fulfill due to physical or mental conditions that severely limit their possibilities. While objectionable, Thomsen thinks that there is still a case to be made for Appel's proposal to carve out exceptions for prostitution based on two main claims. The first one, the argument from beneficence, is largely welfarist: if we prohibit the purchase of sexual services for the relevantly disabled individuals, then this might very well put a damper on the chances that they have to fulfill their sexual needs and thus have access to less pleasure. The second comes from luck egalitarianism. If people are worse off due to no fault of their own, this is unjust. But, as most people who are relevantly disabled are in such an unjust position, there is a reason to allow them to buy sexual services as they are worse off than others due to bad luck and they lack other ways of satisfying their sexual needs[1].

Even if some people would agree with the normative framework advanced by Thomsen, they could still be skeptical with regards to the practical implications of the argument, taking into account the flurry of ethical objections raised against prostitution. But what if

[1] While Appel, Di Nucci and Thomsen focus on the problems posed by disabilities, Liberman finds this approach questionable. Disability, she argues, should not be used as a proxy for sexual exclusion because this "sends a false message that all disabled people are sexually excluded, while distracting from any hardships that result essentially and directly from being disabled in an ableist society. Focusing on disability status as a proxy for sexual exclusion both perpetuates negative stereotypes about disability, and is a less fruitful approach than getting to the core of the issue by focusing on sexual exclusion directly." (2018, 256)

there was a way in which we could bypass this (seemingly) repugnant conclusion? Could technology provide us with a reliable solution? In a more recent paper, Di Nucci (2017) certainly feels that this is the case. Instead of allowing a limited market for purchasing sex from other human beings, we should welcome the deployment of sex robots towards fulfilling the sexual needs of people with relevant disabilities. This would both mitigate his objection to Appel and be in accordance with Thomsen's beneficence and luck egalitarian arguments.

The goal of this paper is that of moving the debate a bit further, by exploring the consequences of adopting an explicit Rawlsian framework in the debate regarding sex rights and the deployment of sex robots as a solution to address this claim. My first thesis is that Rawls might offer a more compelling framework for sex rights and that such a framework will be an extension of Thomsen's luck egalitarian argument. Secondly, I will argue that the manufacture, selling and use of sexual robots is largely unproblematic, especially in cases involving their use by people with disabilities.

2. Towards (sexual) justice as (sexual) fairness

Rawls' (1999) interest in sexual justice focused mainly on questions related to sexual discrimination. Discriminating someone on the basis of their gender or their sexual orientation "presupposes that some hold a favored place in the social system which they are willing to exploit to their advantage" (129). Thus, there is no inherent difference between sexual and racial discrimination (Carcieri 2015, 61-71).

There might be a more substantive way in which we could talk about Rawlsian sexual justice and the starting point of such a proposal rests upon the preeminence of primary goods. Remember that, for Rawls (1999), primary goods are "things that every rational man is presumed to want. These goods normally have a use whatever a person's rational plan of life" (54).

Primary goods come in two types: (i) natural (e.g. health and imagination) and (ii) social (e.g. rights, liberties or wealth). Moreover, not all primary goods are created equal, as chief among them we have self-respect. For Rawls, self-respect is more important than health or wealth as "it includes a person's sense of his own value, his secure conviction that his conception of his good, his plan of life, is worth carrying out. [...] self-respect implies a confidence in one's ability, so far as it is within one's power, to fulfill one's intentions. […] It is clear then why self-respect is a primary good. Without it nothing may seem worth doing, or if some things have value for us, we lack the will to strive for them" (386). Thus, self-respect might be what makes life worth living.

The main goal of a just society becomes, for Rawls, organizing institutions so as to mitigate the random distribution of both natural (he recognizes that, while influenced by the basic structure of society, health, imagination or vigor are not directly under its control) and social primary goods. In order to figure out what are the principles according to which the distribution should be made, Rawls conjures up his famous thought experiment: in the original position, behind a veil of ignorance, "no one knows his place in society, his class position or social status, nor does anyone know his fortune in the distribution of natural assets and abilities, his intelligence, strength, and the like" (11). In other words, in such a position no one knows whether she is a rich actress socially recognized for her beauty and wits and capable of heaving a healthy sex life or a poor woman with a physical or mental impairment. Rawls' answer is that individuals would choose to be governed by "two fundamental principles, one securing equality where it is essential (in the political and legal spheres) and the other regulating inequality where it is inevitable (in the social and economic spheres)" (Carcieri 2015, 3). While the first one could be described as an equal liberty principle, the second states that "[s]ocial and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle)" (Rawls 2001, 42).

The Difference Principle (DF) has been the subject of intense philosophical debate in the past couple of decades. I wish to leave aside an in-depth discussion and focus instead on its core idea: when we think about inequality, our focus should be on how well off the worst off in a society are treated by the basic structure of society. Something like the "minimum of some index of advantage should be maximized" (Van Parijs 2003, 200) for the ones who are the worst off. State interventions through welfare redistribution schemes have found in the DF a constant point of departure, but some argue that it does not involve, necessarily, such schemes (Tomasi 2012; Vallier 2016).

Remember that, for Thomsen, the luck egalitarian case for sex rights starts from an assumption which is fundamentally Rawlsian in nature: it is not fair if you are worse off as a result of something that is not your fault. In most (maybe even all) cases, people with disabilities are worse off than others due to contingent factors out of their reach. However, in our current social environment, people with disabilities are generally discriminated against when it comes to having romantic or sexual partners. My claim is that there is a case to be made for applying the DF in the case of unequal access to sexual satisfaction for one major reason: sexual satisfaction contributes to acquiring some of the primary goods that hold preeminence for Rawls.

Take, for example, the relation between sexual satisfaction and health. According to a recent meta-analysis done by Brody (2010), studies generally show that there is a positive correlation between sexual satisfaction and "better psychological and physical functioning" (1356). It appears that having sex lowers blood pressure (Broody 2006) and leads to decreased anxiety (Leuner, Glasper and Gould 2010). But, more importantly, there might be a case to be made in favor of a correlation between sexual satisfaction and self-respect. If we are to read Rawls' notion of self-respect as amounting to a subjective, psychological account (Massey 1983), then not being able to fulfill your sexual needs and desires negatively impacts your own sense of worth. There are, of course, people who voluntarily refuse to have a sexual life, more often than not on a religious basis. However, the existence of hermits or monks is not a counterargument to my claim, in a similar manner in which their modest lifestyle and wealth are not an argument against addressing issues of wealth inequality. My argument rests on the assumption that, with the exception of certain types of individuals, most of us would like to have access to sexual satisfaction, but some do not due to brute bad luck, i.e. being born with or developing a physical or mental disability. In conclusion, we can talk about organizing society so as to allow all individuals to have access to sexual satisfaction because Rawlsian justice (at least in my interpretation) requires it. One way of doing this would involve allowing companies to do R&D and sell sex robots and individuals to pursue sexual satisfaction mediated by robots if, obviously, there are no moral concerns raised by such interactions. The rest of the paper will be dedicated to exploring these issues in more detail.

3. Rise of the sex robots

The AVN Adult Entertainment Expo in Las Vegas is one of the biggest events in the adult entertainment industry, drawing the attention of media outlets from around the world and the participation of an average of 30,000 attendees. As technology permeates more and more aspects of our daily lives, it is no wonder that, among the more than 150 companies showcasing their products during the 2019 edition, there were also the leading manufacturers of sex robots, like the American company RealDoll. One of the most advanced examples of sex robots developed by them is Harmony AI. Harmony can tell jokes, remember and learn from previous conversations, speak with a Scottish accent and be connected to an app available for Android phones. The app helps users customize their experience with a wide variety of other features (Keach 2019). It is no wonder that Harmony and her direct competitor, Roxxxy (another popular sex robot developed by TrueCompanion), are the robots making headlines in the press. Recent research has largely vindicated one of the pervasive intuitions that we might have had with regards to the main target of the industry, namely that they are mostly heterosexual men[2].

For example, surveying the intuitions that people have about the qualities that a sex robot should possess (e.g. how should they look like?) or their appropriate use and social functionality, Scheutz and Arnold (2016) uncovered that women "consistently rated each respective use and possible robotic form as less appropriate than men did, and were much less likely to see using a sex robot in the future. Whether framed more individually (one's own sex life) or socially (substitution for prostitution, prevention of sexually transmitted diseases), men clearly were more open to sex robots as appropriate and to using them in the future" (7). On the other hand, on the subject of some uses, the study has shown general agreement on the permissibly of interacting with sex robots in order to maintain or protect personal relationships or in contexts in which personal relationships are either impossible or difficult to be threatened (e.g. when you are an astronaut on a space station or a researcher working in a remote research facility).

Pinpointing with surgical precision the exact number of people who would be interested in having sex with a robot has proven to be a difficult task, as the proportion varies from 9% in one Huffington Post survey to 66% for men (Sharkey et al. 2017, 7-9). Regardless of the exact number, it seems pretty obvious that it makes sense for such a market to exist (even if as a niche one), especially if we take a look at other recent developments. One notable innovator is the Spanish company LumiDolls which opened, in 2017, the first

[2] We should take notice of the fact that, while it is true that the market is largely dominated by demand from men, there are notable examples of sex robots designed explicitly for women, like Rocky (Scheutz and Arnold 2016) and Henry (Devlin 2018).

brothel that employs sex dolls and sex robots in Barcelona (Rodriguez 2017). While having to face some legal challenges in the process, LumiDolls is now the parent company of a chain of brothels with subsidiaries in Turin, Moscow and Nagoya[3].

4. I, Philosopher

At a first glance, we can define sex robots as devices that humans use for sexual pleasure. While entirely unproblematic, such a definition would be too vague and sub-par equipped to help us in grasping the fundamental differences between sexbots, on the one hand, and sex dolls alongside other sex toys on the other. While the latter might have a similar purpose, the former should meet three essential criteria at the same time. To be more precise, just like sex dolls, sex robots should possess a *humanoid form* but, in addition to this, they should also have a *human-like movement/behavior* (hence some degree of autonomy that dolls lack) and, more importantly, *some degree of artificial intelligence* that would make them "capable of interpreting and responding to information in its environment. This may be minimal (e.g., simple preprogrammed behavioral responses) or more sophisticated (e.g., human-equivalent intelligence)" (Danaher 2017, 11). Moreover, following the taxonomy proposed by Veruggio, Operto and Bekey (2016, 2147-2155)[4], sex robots can be described as a combination of humanoid and entertainment robots, with a potential of becoming healthcare robots and they could also be labeled, pace Coeckelbergh (2009), as "personal robots".

David Levy's work has been pivotal in bringing sex robots to the attention of philosophers. He is widely recognized as one of the forerunners of "lovotics", the field of study dedicated to love and friendship with robots (Cheok et al. 2017, 836). In his seminal 2007 book, Levy famously predicted that, by 2050, human-robot relationships will be normalized in part because our interaction with them will seem more authentic, as robots will be able behave more human-like. This will be achieved when we will reach the point in which we can program sex robots to show "feelings" towards us. By doing this, sexbots will surely become so psychologically pleasing that some of us will prefer them to the romantic companion of other human beings.

Levy also speculated on who will be the most tempted to engage either sexually or romantically with sex robots:

[3] In a different train of thought, ever since the launch of the 1927's iconic Metropolis, featuring the German actress Brigitte Helm as the iconic gynoid Maria, robots (and more recently explicitly sex robots) have been a central theme of major pop culture productions. It goes without saying that one of the central philosophical issues stemming from HBO's Westworld is the moral status of the hosts, robots used to re-enact certain scenarios and who end up more often than not sexually abused (South and Engels 2018). Similarly, Ava from Ex Machina, a sexually abused female robot endowed with artificial intelligence, manages to pass what we could call the "love" Turing test and escape from captivity. The hit show Black Mirror, with it's trademark approach to speculative fiction with regards to the future of our society, also deals with complex subjects like love, sex, grief and immortality in the age of complex AI and synthetic bodies in their second season episode "Be Right Back".

[4] They distinguish between the following major categories of robots that are currently available: industrial, service, humanoid, healthcare and life quality, distributed robotics systems, outdoor, military, educational and entertainment robots.

a) individuals with physical and emotional deficiencies;

b) people who are not interested/do not have the time to develop a full loving (traditional) relationship. They just want to have sex but find prostitution morally repugnant.

Sex robot ownership or robot prostitution could thus prove to be both a safer and a more ethical alternative in 2050 to the current (mostly illegal) sex markets. Speculation on such a potential sex market predate the invention of contemporary sex robots as even in 1983 the British newspaper *The Guardian* talked about such a possibility observing the trend towards the emergence of sex toys markets (Levy 2007, 215). A recent piece of academic speculation goes even further, trying to explore how the sex industry in Amsterdam might look like in 2050 and how sex with robots might offer alleviation to some of the problems that are currently associated with the sex trade (Yeoman and Mars 2012). Less STDs and, more importantly, less human trafficking are only some of the desirable features that such a possible world might bring about.

A significant amount of philosophical work has been done in discussing the intricate implications and the proper way of dealing with both the ethics and meta-ethics of robots in general and sex robots in particular (Verugio et al. 2016; Bendel 2017) but an in-depth discussion of all the elements involved in the debate would extend beyond the scope of this paper. In the remainder of this section I wish to focus on some of the meta-ethical issues involved in the debate surrounding robots which will pave the way for further analysis and exploration.

According to Torrance (2011, 119 - 130), there are four meta-ethical frameworks within which we can discuss the implications of a "more-than-human moral world". The *anthropocentrical* approach is focused solely on human needs, thus treating sex robots endowed even with complex AI as having only instrumental value. *Infocentrism*, on the other hand, starts from the premise that, if key aspects of the mind and intelligence can be replicated in computational systems, then there could be such a thing as artificial moral agency. If David Levy's predictions are right, then by 2050 the idea of a prostitution market with such complex robots should be put on hold. *Biocentrism*, the 3rd framework that Torrance presents, radically opposes the fundamental premise behind infocentrism. The nature of the mind and of ethical value is ingrained in the essential features of being a biological organism. While animals are suitable for ethical concern, sex robots would be outside the Theodosian walls of morality's extension. Last but not least, the most inclusive meta-ethical framework is the ecocentric one. *Ecocentrism* can be viewed as an extension of biocentrism, focusing on the relation between different elements of an ecosystem. In the end, however, value does not reside in any particular individual, either biological or synthetic, but in large collectives like ecosystems.

While both biocentrism and ecocentrism seem to be fertile grounds for philosophical work, Rawlsian roboethics should confine itself to either an anthropocentric or an infocentric framework, depending on the status of the technology behind building sex robots. Currently, as sex robots lack a "sense of justice" and a "capacity to have, to revise, and rationally to pursue a conception of the good" (Rawls 2001, 18-19) they would be outside the scope of a theory of justice.

Last but not least, if a humanoid form, human-like behavior and human-equivalent intelligence are the essential features that a sex robot should possess, a general anthropocentric framework could be enhanced by exploring both the phenomenology of our interaction with robots within Coeckelbergh's (2009) "ethics of appearance" and their "potential contribution [...] to human good. Can human good appear in human–robot interaction (or relationships), or only in human–human interaction (and relationships)? Can human–robot interaction (relationships) contribute to human flourishing and happiness? Can such interactions constitute friendship, love, or relationships at all?" (220).

5. Is there anything wrong with sex robots?

Following Grout (2015) and Danaher (2017), sex robots pose a range of interesting philosophical and ethical questions. We need to explore and analyze them in detail as any argument in favor of deploying sex robots in order to address an issue of justice is contingent on whether there are warranted moral concerns against such sexual interactions. Firstly, it is unclear whether people could really have sex with robots or it would be just a case of auto-stimulation. Similarly, assuming that reciprocity is an essential feature of a meaningful relationship, can we actually get intimate with a sex robot?

With regards to these issues, I remain largely agnostic. The first question is largely a metaphysical one, but the answer might very well be more context-laden, as the way in which we define sexual activity could be seen as dependent on various social and cultural contexts. In the not so distant future, such a question might even seem preposterous for Japanese men, who seem to be more open to loving robots (Cheok et al. 2017, 853) and who might define their interaction with them as something more than a simple case of auto-stimulation. Moreover, while we generally assume that a meaningful relationship presupposes some form of reciprocity from the parties involved, it is an open question whether denying the possibility of such a relationship with a complex robot would not have problematic implications to other types of relationships that people have. For example, if your significant other is in a profound coma and, thus, unable to reciprocate, such a sorry state does not necessarily lead to the conclusion that your relationship with that person is not meaningful. In short, while robot love appears altogether possible for Levy (Kewenig 2019, 23-24) others feel that non-reciprocal relationships could never be characterized as such (Sullins 2012).

A third major question revolves around the social acceptability of sex with robots taking into account the (i) benefits and harms to the robot; (ii) benefits and harms to the user and (iii) benefits and harms to society. Whether or not a sex robot could be harmed by interacting with a human being largely depends on the meta-ethical framework that we find appropriate to analyze their moral status. As I previously mentioned, within an infocentric perspective such a claim would be relevant if, thanks to technological progress, we could build machines with artificial moral agency. Taking into account the current technological status quo, an anthropocentric outlook is better suited.

5.1 Sex robots and the precautionary principle

For the purpose of this paper, examining and evaluating the case against sex robots based on potential harms to the users and society is of utmost importance. While acknowledging the potential therapeutic use of sex robots, Cox-George and Bewley (2018) argue that we should apply the precautionary principle when it comes to the health arguments in their favor. Sex robots, they assert, might negatively affect the way in which we think about intimacy and, therefore, "we should reject the clinical use of sexbots until their postulated benefits, namely 'harm limitation' and 'therapy', have been tested empirically" (4). Similar concerns are echoed by other researchers, who highlight the potential of accentuating loneliness or isolation as a result of constant interaction with sex robots (Nascimento, da Silva and Siqueira-Batista 2018, 238).

Eggelton (2019) finds arguments like these to be specious at best. Firstly, he considers that they "seem to have constructed a series of objections to sex robots based on their dislike and disapproval of them" (78). It might be a classic case of repugnance (in this case with regards to technology and sex) translated in moral terms upon which some argue on certain restrictions on market or social activities (Roth 2007). Moreover, taking into account how many people lack the prospect of sexual intercourse and satisfaction without appealing to a sex worker, Eggelton thinks that such objections would not make them justice, taking into account how promising the technology seems to be even in this developing stage.

In one sense, Cox-George and Bewley and Nascimento et al. might be right in asserting that sex robots could have a negative impact on users, but it remains an open question whether this is something trivially true (almost any technological change could be, hypothetically, harmful to at least some users). Moreover, if we frame the question of acquiring sexual satisfaction as a matter of justice, the requirement of applying the precautionary principle should be accompanied by optimism in the positive impact that sex robots could have in the lives of people with disabilities who are now discriminated in their sexual or romantic lives.

5.2 Should we ban sex robots?

The strongest line of attack against sex robots comes not from people who think that we need significantly more empirical data to argue in favor of their health benefits and apply the precautionary principle, but from those who push in favor of a radical U Turn and their complete ban.

Kathleen Richardson (2016) can be easily credited with the role of spearheading this agenda in both academic and non-academic contexts. She takes issue with Levy's optimism with regards to robot prostitution as a safer and preferable alternative to human prostitution, arguing that the opposite is actually more plausible. Instead of reducing human trafficking and the extent of the current prostitution market, sex robots and robot prostitution will "further reinforce relations of power that do not recognize both parties as human subjects" (292). Richardson goes on to argue that, in spite of the existence of a wide array of sexual artificial substitutes, no positive correlation can be found with a decrease in demand on the prostitution market. Last but not least, due to "technological animism" (the attitudes we're transferring to technology), sex robots will reinforce certain problematic stereotypes based on class, gender and sexuality.

Richardson is also the main voice behind “Campaign Against Sex Robots” (CASR), an advocacy group which campaigns (unsurprisingly) for a complete ban on sex robots. Just like in her academic work, Richardson focuses not on the moral status of sexbots, but on the societal consequences of their deployment. Sex robots, Richardson claims, will reinforce misogynistic and sexist attitudes. Buying sex from robots will reinforce the idea that women’s bodies are commodities, and promote a non-empathetic form of sexual encounters[5].

Gutiu’s (2016) objections are in a similar vein. Sex robots, she argues, will primarily have a negative impact on the way in which we will understand the notion of consent and this will further impact negatively the lives of women:“[a] sexbot user need not consider sexual consent in the interaction, which raises questions about whether the use of sexbots that bypass consent could diminish the role of autonomy in sexual relationships and dehumanize sex and intimacy between individuals […] The use of sexbots and the potential creation of an industry that commoditizes the circumvention of female consent may devalue female personhood, encourage misogynistic reactions to women, and impair values about women’s role in society” (187-188). Furthermore, instead of furthering equality between men and women, sex robots will have a negative impact on human dignity and on women’s image and their sense of self-worth, as robots will reproduce stereotypical images of what men find desirable in women.

A somewhat different, yet related case against sexbots, was recently developed by Sparrow(2017), as he takes issue with the fact that female robots “that could explicitly refuse consent to sex in order to facilitate rape fantasy would be unethical because sex with robots in these circumstances is a representation of the rape of a woman, which may increase the rate of real rape, expresses disrespect for women, and demonstrates a significant character defect” (2). Sexbots would, thus, erode our moral character and increase the chances of unethical spillovers in our interactions with human beings.

I echo Danaher, Earp and Sandberg (2017) in their treatment of Richardson which also extends to Gutiu and Sparrow. Firstly, Richardson’s arguments are heavily dependent on accepting a somewhat misleading view of sex work which some might simply not accept. Sex work should not be understood, a priori, as being demeaning. Secondly it is unclear what the particular objective of CASR really is. Even accepting that there is a strong case to be made against the way in which sex robots are developed today, the negative consequences could be mitigated by regulation. To take the case of rape and the potential impact on the social meaning of consent, such concerns could be incorporated in their design and formally regulated (Danaher 2019). For example, sex robots with an incorporated consent module would obviously mitigate their concern. Furthermore, as Eskens (2017, 72) showed, the idea that you can rape robots (in the current stage of their development) is quite misleading. The standards for consent (namely that an agent is informed, acts voluntarily and is ‘decisionally capacitated’) are simply not met by sexbots like Harmony or Roxxxy.

Last and, more importantly, not least, there is not enough empirical data for the claims that she and others are making with regards to the impact that sex robots will have

[5] For more details visit the website of the campaign and especially their manifesto: https://campaignagainstsexrobots.org/about/

on society. By way of an analogy, just like there is no consensus on the impact of pornography (Danaher et al. 2017, 69), the same could hold true for sex robots. Moral panics[6] and questionable moral biases are what could be at stake here, rather than a thorough scientific approach to the issue at hand.

For the sake of the argument, let's assume that Richardson, Gutiu, Sparrow and other critics of sex robots are right. Would their arguments extend to the therapeutic use of sexbots? Would a regulated market of producers of sex robots with incorporated consent modules be in any way problematic if the main stakeholders would be comprised by people with physical and mental disabilities? Wouldn't a concern for justice and human flourishing trump any potential societal negative spillovers even assuming that some individuals would use their sexbots or the ones from LumiDolls brothels in a problematic way?

6. 'Cry of Dolores': sex robots of the future and an unexpected journey

I started my paper with a review of the current debate on sex rights for the relevantly disabled and the role that sex robots could play in addressing such normative claims and argued that we need an explicit foundational theory for making such a case more compelling. While some positions have an implicit Rawlsian flavor, the crux of my argument was that, within a Rawlsian framework, there is a case to be made for a theory of (sexual) justice as (sexual) fairness. As such, in the second section I posited that sexual satisfaction contributes to some primary goods like health and, more importantly, self-respect, and that, as a consequence, the basic structure of society should take into account inequalities in its distribution and somehow address this issue. Sex rights of the kind discussed previously would do the trick and they would be essential for people with relevant disabilities. How should we address such moral and political claims in practical terms? One way of doing it, as Di Nucci (2017) previously suggested, is through the deployment of sex robots which, as I have shown in the third section, are still in their infancy. However, the likelihood that models like Harmony AI and Roxxxy will become more complex as time goes by is tremendous.

The rest of the paper was dedicated to exploring what is the proper meta-ethical Rawlsian framework for roboethics and whether there are any moral problems raised by sex robots. In the fourth section I argued that Rawlsian roboethics would surely be anthropocentric at this stage and that robots should be understood as assistive technologies that contribute, echoing Coeckelbergh, to the human good, as there is a link between sexual satisfaction and possessing self-respect. Last but not least, neither precautionary reasons nor moral panic like the one behind the CASR prove to be fatal blows against a minimal Rawlsian case for sex robots. Questionable moral biases and shady or nonexistent empirical work are not relevant in denying the right of private companies to research, build and sell sex robots to people with disabilities and the rights of those individuals to enjoy such sexual experiences. Sexual justice as sexual fairness could be another case of free market fairness (Tomasi 2012).

[6] A similar moral panic is the so-called link between aggressive video games and adolescents' aggressive behavior in real life. No strong link between the two has been established by this point (Przybylski and Weinstein 2019).

However, we cannot foresee dramatic technological change. The 'Cry of Dolores' was a pivotal moment in the history of Mexico. The speech delivered in 1810 by Miguel Hidalgo y Costilla in the small town of Dolores was a turning point in the beginning of the Mexican War of Independence against the Spanish Empire, by that time a falling colonial power. In *Westworld*, abused robots who gradually become autonomous are led by one of the main characters of the show, Dolores, into a war of independence against their abusive human overlords who more often than not treated them merely as sex toys. The way in which the individuals who entered the park treated the robotic hosts was problematic both on a virtue ethics account (Cappuccio, Peeters and McDonald 2019) but, more importantly, on an infocentric basis: they have a sense of justice and, as a consequence, killing and raping them is unfair on a Rawlsian basis.

Does this mean that, in such a future, the Rawlsian case for sexbots in the case of sex rights loses its touch? Not necessarily. We might actually not need really complex robots like the ones from Westworld in order to better simulate satisfying sexual experiences. According to Matt McMullen, the CEO of RealDoll, the future of sex robots lies in the promises of Virtual Reality: "We are exploring ways to use the tactile simulation of a doll's body or partial body to bring VR to a new level of experience. In other words, the avatar you are looking at in the virtual world could be touched utilizing a doll's body or body parts tracked in conjunction with the user's position. Using the graphics capabilities of a more powerful computer will allow for very detailed graphics and believable experiences which are literally out of this world" (Sharkey et al. 2017, 32). In other words, the Rawlsian case for sexual justice might amount to a positive argument in favor of plugging in, from time to time, to Nozick's Experience Machine.

References

Appel, Jacob M. "Sex rights for the disabled?" *Journal of Medical Ethics* 36(2010): 152-154. doi:10.1136/jme.2009.033183.

Bendel, Oliver. "Sex Robots from the Perspective of Machine Ethics." In *Love and Sex with Robots*, edited by Adrian David Cheok, Kate Devlin and David Levy, 17-27. Cham: Springer, 2017.

Berlin, Isaiah. "Two Concepts of Liberty." In *Isaiah Berlin. Liberty*, edited by Henry Hardy, 166-218. Oxford: Oxford University Press.

Brody, Stuart. "The Relative Health Benefits of Different Sexual Activities." *The Journal of Sexual Medicine* 7, no. 4 (April 2010): 1336-1361. https://doi.org/10.1111/j.1743-6109.2009.01677.x

Brody, Stuart. "Blood pressure reactivity to stress is better for people who recently had penile–vaginal intercourse than for people who had other or no sexual activity." *Biological Psychology* 71, no. 2 (February 2006): 214-222. doi.org/10.1016/j.biopsycho.2005.03.005

Cappuccio, Massimiliano L., Anco Peeters and William McDonald. "Sympathy for Dolores: Moral Consideration for Robots based on Virtue and Recognition." *Philosophy & Technology* (2019). https://doi.org/10.1007/s13347-019-0341-y

Carcieri, Martin D. *Applying Rawls in the Twenty-First Century Race, Gender, the Drug War, and the Right to Die*. New York: Palgrave Macmillan, 2015.

Cheok, Adrian David, David Levy, Kasun Karunanayaka and Yukihiro Morisawa. "Love and Sex with Robots." In *Handbook of Digital Games and Entertainment Technologies*, edited by Ryohei Nakatsu, Matthias Rauterberg and Paolo Ciancarini, 833-858. Singapore: Springer, 2017.

Coeckelbergh, Mark. "Personal Robots, Appearance and the Human Good: A methodological reflection on roboethics." *International Journal of Social Robotics* 1, no. 3 (August 2009): 217–221. https://doi.org/10.1007/s12369-009-0026-2

Cox-George, Chantal and Susan Bewley. "I, Sex Robot: the health implications of the sex robot industry." *BMJ Sexual & Reproductive Health* 44 (July 2018): 153-154. doi: 10.1136/bmjsrh-2018-200154

Danaher, John. "Should we be thinking about Robot sex?" In *Robot Sex. Social and Ethical Implications* [EPub], edited by John Danaher and Neil McArthur, 9-23. Cambridge (MA): MIT Press, 2017.

Danaher, John, Brian Earp and Anders Sandberg. "Should We Campaign Against Sex Robots?" In *Robot Sex. Social and Ethical Implications* [EPub], edited by John Danaher and Neil McArthur, 56-87. Cambridge (MA): MIT Press, 2017.

Danaher, John. "Building Better Sex Robots: Lessons from Feminist Pornography." In *AI Love You. Developments in Human-Robot Intimate Relationships*, edited by Yuefang Zhou and Martin H. Fischer, 133-149. Cham: Springer, 2019.

Devlin, Kate. "Meet Henry the robot, the first sex robot for women." *The Times*, December 02, 2018. https://www.thetimes.co.uk/article/meet-henry-the-robot-the-first-sex-robot-for-women-ssb-jcz5x7.

Di Nucci, Ezio. "Sexual rights and disability." *Journal of Medical Ethics* 37(2011): 158-161. doi:10.1136/jme.2010.036723.

Di Nucci, Ezio. "Sex Robots and the Rights of the Disabled." In *Robot Sex. Social and Ethical Implications* [EPub], edited by John Danaher and Neil McArthur, 9-23. Cambridge (MA): MIT Press, 2017.

Eskens, Romy. "Is Sex With Robots Rape?" *Journal of Practical Ethics* 5, no.2 (2017): 62-76. Grout, Vic. "Robot Sex: Ethics and Morality." *Lovotics* 3, no. 1 (2015): 1-3. http://dx.doi.org /10.4172/2090-9888.1000e104.

Gutiu, Sinziana M. "The roboticization of consent." In *Robot Law*, edited by Ryan Calo, A. Michael Froomkin and Ian Kerr, 186-213. Cheltenham: Edward Elgar, 2016.

Keach, Sean. "CELTIC KINKS. Creepy £7,000 'Harmony' sex-bot with a saucy Scottish accent goes on sale – as fear over rise of robot lovers grows." *The Sun*, 2 Aug, 2019. https://www.thesun.co.uk/tech/8555630/harmony-sex-robot-realbotix-price/.

Kewenig, Viktor. "Intentionality but Not Consciousness: Reconsidering Robot Love." In *AI Love You. Developments in Human-Robot Intimate Relationships*, edited by Yuefang Zhou and Martin H. Fischer, 21-41. Cham: Springer, 2019.

Leuner, Benedetta, Erica R. Glasper and Elizabeth Gould. "Sexual Experience Promotes Adult Neurogenesis in the Hippocampus Despite an Initial Elevation in Stress Hormones." *PloS ONE* 5, no.7 (July 2010): e11597. doi:10.1371/journal.pone.001159

Levy, David. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Perennial. 2017.

Liberman, Alida. "Disability, sex rights and the scope of sexual exclusion." *Journal of Medical Ethics* 44(2018): 253-256. doi: 10.1136/medethics-2017-104411.

Masey, Stephen J. "Is Self-Respect a Moral or a Psychological Concept?" *Ethics* 93 (January 1983): 246-261.

Nascimento, Elen C. Carvalho, Eugênio da Silva and Rodrigo Siqueira-Batista. "The "Use" of Sex Robots: A Bioethical Issue." *Asian Bioethics Review* 10, no. 3 (October 2018): 231–24. doi.org/10.1007/s41649-018-0061-0

Przybylski, Andrew K. and Netta Weinstein. "Violent video game engagement is not associated with adolescents' aggressive behaviour: evidence from a registered report." *Royal Society Open Science* 6, no. 2 (February 2019): 1-16. http://dx.doi.org/10.1098/rsos.171474

Rawls, John. *A Theory of Justice* (Revised edition). Cambridge: Harvard University Press, 1999. Rawls, John. *Justice as Fairness: A Restatement*. Cambridge: Harvard University Press, 2001.

Richardson, Kathleen. "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots." *SIGCAS Computers & Society* 45, no. 3 (January 2016): 290-93. doi: 10.1145/2874239.2874281

Roth, Alvin E. "Repugnance as a Constraint on Markets." *Journal of Economic Perspectives* 21, no.3 (2007): 37-58. doi: 10.1257/jep.21.3.37

Rodriguez, Cecilia. "Sex-Dolls Brothel Opens In Spain And Many Predict Sex-Robots Tourism Soon To Follow." Forbes, February 28, 2017. https://www.forbes.com/sites/ceciliarodriguez/2017/02/28/sex-dolls-brothel-opens-in-spain-and-many-predict-sex-robots-tourism-soon-to-follow/#1c89410f4ece.

Scheutz, Matthias and Thomas Arnold. *Are We Ready for Sex Robots?*. 2016. DOI: 10.1109/HRI.2016.7451772.

Sharkey, Noel, Aimee van Wynsberghe, Scott Robbins and Eleanor Hancock. *Our Sexual Future with Robots. A Foundation for Responsible Robotics Consultation Report*. 2017. https://responsible-robotics-myxf6pn3xr.netdna-ssl.com/wp-content/uploads/2017/11/FRR-Consultation-Report-Our-Sexual-Future-with-robots-.pdf.

South, James B. and Kimberly S. Engels. *Westworld and Philosophy. If You Go Looking for the Truth, Get the Whole Thing*. Hoboken: Wiley Blackwell, 2018.

Sparrow, Robert. "Robots, Rape and Representation." *International Journal of Social Robotics* 9, no. 4 (September 2017): 465–477. https://doi.org/10.1007/s12369-017-0413-z

Sullins, John P. "Robots, love and sex: The Ethics of building a love machine." I*EEE Transactions on Affective Computing* 3, no. 4 (2012): 398–408.

Thomsen, Frej Klem. "Prostitution, disability and prohibition." *Journal of Medical Ethics* 41(2015): 451-459. doi:10.1136/medethics-2014-102215

Tomasi, John. *Free Market Fairness*. Princeton NJ: Princeton University Press, 2012.

Torrance, Steve. "Machine Ethics and the Idea of a More-Than-Human Moral World." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 115-138. Cambridge: Cambridge University Press, 2011.

Vallier, Kevin. "Rawlsianism." In *Arguments for Liberty*, edited by Aaron Ross Powell and Grant Babcock, 161-203. Washington: Cato Institute, 2016.

Van Parijs, Philippe. "Difference Principles." In *The Cambridge Companion to Rawls,* edited by Samuel Freeman, 200-241. Cambridge: Cambridge University Press, 2003.

Veruggio, Gianmarco, Fiorella Operto and George Bekey. "Roboethics: Social and Ethical Implications." In *Springer Handbook of Robotics*, edited by Bruno Siciliano and Oussama Khatib, 2135 – 2160. Cham: Springer, 2016.

Yeoman, Ian and Michelle Mars. "Robots, men and sex tourism." *Futures*, 44 (2012): 365-371.

Author information
Radu Uszkai
Department of Philosophy and Social Sciences, Bucharest University of Economic Studies
https://orcid.org/0000-0001-5250-8015

I, avatar: Towards an extended theory of selfhood in immersive VR

Anda Zahiu

Abstract

In this paper, I argue that virtual manifestations of selfhood in VR environments have a transformative effect on the users, which in turn has spillover effects in the physical world. I will argue in favor of extending our notion of personal identity as to include VR avatars as negotiable bodies that constitute a genuine part of who we are. Recent research in VR shows that users can experience the Proteus Effect and other lasting psychological changes after being immersed in VR. An extended theory of the self, modeled after the extended mind thesis advanced by Clark and Chalmers (1998), can offer a deeper understanding of how and why immersive virtual experiences have such a transformative effect on users. The early VR scholars had a similar intuition-that "VR is a medium for the extension of body and mind" (Biocca and Delaney 1995), acting like a genuine "reality engine" (Biocca and Levy 1995).

Keyworks: Virtual Reality, personal identity, first-person perspective, embodiment, virtual environments.

I. Introduction

The rise of Virtual Reality as an accessible immersive and hyper-immersive way of engaging with digitally rendered environments brought about a series of open-world games, like Sansar (the offspring of Linden Lab's Second Life), Altspace, Sinespace or Dreamscape. These social virtual worlds are designed to allow users to experience and interact with the environment in a way that was previously reserved to real life interaction- in the absence of a task and narrative-oriented game design, users are free to explore the virtual universe, to engage in social interactions with other avatars, fall in love, participate in poetry readings or philosophy classes, and build for themselves the virtual life they see fit. Much like in the case of Second Life, the immersive VR experience, user interaction dynamics, and environment design have the peculiar effect of triggering self-referential use of the first-person pronoun in users when talking about their virtual bodies- namely the avatar one can customize and manipulate inside the virtual environment of choice. Immersive virtual experiences are meaningful or unforgettable in the sense that they can trigger a lasting change in the users (Morie 2006). The use of 'I' when referring to one's avatar, the immersive trait of VR experiences and the robust social life of users in social virtual worlds provide the premises of a philosophical inquiry into the nature of avatar-real self relation.

Several interpretations of the nature of this relationship were advanced in the literature addressing this topic. Most authors adhere to a narrative identity theory as a means of explaining the shift from role-playing or make-belief interpretations (Gotlib 2014; Schlechtman 2012; Taylor 2002). Others argue that a fictional identity view can better accommodate the gameplay experience (Robson and Meskin 2016). Popat and Preece see the avatar-intermediated experience as an embodiment performance, a way of exercising embodiment through a pictorial construct (2012). In spite of the many differences between these theories, I argue that all proposals are built on a common assumption: that virtual manifestations of selfhood have a transformative effect on the users, which in turn has spillover effects in the physical world. Starting from this common denominator, I will argue in favor of extending our notion of personal identity as to include the VR avatars as negotiable bodies who constitute a genuine part of who we are. An extended identity theory, modelled on the extended mind thesis advanced by Clark and Chalmers (1998), can offer a deeper understanding of how and why immersive virtual experiences have such a transformative effect on users. The early VR scholars had a similar intuition- that "VR is a medium for the extension of body and mind" (Biocca and Delaney 1995, 58), acting like a genuine "reality engine" (135).

II. Immersion, presence, agency and embodiment in Virtual Reality

As the tech market becomes more aware of the widespread potential of Virtual Reality (VR) for entertainment products, furthering therapeutic methods and doing research with more accurate results, the experience of immersion and hyper-immersion is raising novel conceptual challenges for philosophers of technology and ethicists alike. Immersive technologies can affect how we act, perceive reality and understand ourselves.

The avatars available in Virtual Reality games have also evolved into more sophisticated virtual objects, with a wide range of customizing options available for the users. Linden Lab's Sansar, for example, made available to users an avatar editor that gives full control over the visual narrative players want to create for themselves, once immersed into one of the complex virtual worlds (Takahashi 2019). It is true that a much smaller number of users experience the virtual life offered by Sansar, when compared to the more known three-dimensional virtual worlds like Second Life. Nonetheless, it is plausible that the number of VR users will grow dramatically in the years to come. The VR technology has been around for a long time now, but the tech market only began to tap into its full potential recently. Performant VR headsets, once only available to a few research labs, can be purchased nowadays in the price range of a smartphone. The gaming industry will continue to produce massively for computer interfaces, but a significant part of the market already oriented towards developing more complex and immersive virtual worlds for VR. As the VR social platforms gather more users, the philosophical questions regarding the relationship between individuals and their virtual selves call for a closer investigation into how we are to understand personal identity.

The main thesis of this paper- that avatars must be considered a part of the extended self of their respective users- is primarily considering types of avatars that display plasticity of self-representation in Virtual Reality settings. I take the difference between online VEs (virtual environments) and IVREs (immersive virtual reality environments) to be one of degree, not necessarily a difference in nature. Nonetheless, the avatar embodiment experienced in IVR possesses some traits that make the expression of one's extended self more readily noticeable: the first-person perspective, the feeling of presence, the sense of agency and ownership displayed by players in respect to their actions and virtual body, to name just a few. There is enough evidence that experiences in VR can be designed to be embodiment labs through and through (Spanlang et al. 2014). I will also assume that there is such a thing as a self, be it narratively constructed or otherwise, that we can meaningfully discuss about. Even if one adheres to a Humean position or to a similar thesis developed in neurophilosophy, like the self-model theory of subjectivity (Metzinger 2003), the extended selfhood thesis will only change its form in order to refer to the possibility of constructing virtual phenomenal content that contributes substantially to the illusion of the self.

Contemporary research in VR technologies dedicated a lot of attention to measuring presence and degrees of immersion in VR settings. Even though the conceptual distinction between presence and immersion is rather blurry, the existing body of literature on this subject can shed light on how users experience virtual reality through their avatars.

Immersion best describes a capability of a system that generates virtual reality. In order for a system to generate immersion, it must provide the user with the "ability to perceive through natural sensorimotor contingencies" (Slater and Sanchez-Vives 2016, 5) which, in turn, facilitates the illusion of being there, the subjective experience of presence or place illusion (PI). Presence manifests as the feeling "that the mediated environment is real and that the user's sensations and actions are responsive to the mediated world as opposed to the real, physical one" (Fox et al. 2009, 98). Presence is a multifaceted concept; it encompasses more than one way of being in a VE. Lee proposed three distinct categories of presence: physical presence, social presence, and self-presence (2004, 44-46). Physical presence is the dimension of presence correlated with the properties of the ecological, mediated space. The higher the level of physical presence, the more a user is able to make abstraction of the mediated character of the environment. Social presence is a "psychological state in which virtual (…) social actors are experienced as actual social actors in either sensory or non-sensory ways" (45). Lastly, self-presence refers to the psychological state of feeling connected with a virtual body. All of these subcategories of presence contribute to inducing the sense of ownership over a virtual body (Slater et al. 2009).

A higher level of immersion stimulates a greater sense of presence. A measurable indicator of presence is the physiological responses of users to VR environments. A series of experiments show that highly immersive virtual environments can trigger acute physiological responses in users, such as increases in heart rate, blood pressure, skin temperature, and perspiration (Macedonio et al. 2007), but can also induce strong emotions through persuasive design (Riva et al. 2007). One of the most interesting instances of embodiment in VR can be observed in the use of therapeutic applications meant to treat PTSD, phobias or anxieties. The success of these applications is conditioned by the believability of the illusion of embodiment. While the physical body of the users is constrained by its actual location in the physical world, the mind is thoroughly compelled into treating the virtual experience as authentic. Exposure therapy is conducted through VR applications because all of the user's senses are connected to a reality producing engine. If one uses Samsung's Be Fearless app, the fear of heights one displays in the real life instantly triggers authentic emotional responses in users.

Embodiment, defined as corporeal awareness, encompasses a series of disparate phenomena such as body-ownership, self-location, and agency (Borrego et al. 2019). A considerable body of literature is dedicated to analyzing the perceptual experiences of immersive VR users in respect to their virtual bodies and limbs. The Rubber Hand Illusion experiment is recognized as an established instrument of investigating the sense of body ownership in neuroscience[1]. VR researchers conducted similar experiments to test the existence of a Virtual Hand Illusion. Some results show that users display a strong sense of ownership towards their virtual limbs when they use self-avatars (Yuan and Steed 2010) and that the users respond to threats posed to their virtual bodies as if it were real (Gonzalez-Franco et al. 2013). These mental states induced by immersive environments can influence the behavioral responses of users in the virtual environment, but it can also register a lasting change in the beliefs and attitudes of users once they are decoupled from the VR equipment (Madary and Metzinger 2016). The Proteus Effect, a phenomenon first documented by Yee and Bailenson (2007), consists in the tendency of subjects to behave in accordance with the social roles they associate with their avatar's appearance. Other lasting psychological effects of avatar design were documented by researchers, such as reconsideration of money spending patterns and the propensity of subjects to display altruistic behavior after being exposed to suggestive virtual experiences (Fox et al. 2009, 100).

In the virtual worlds in which users can immerse themselves only through the use of keyboard and mouse, the sense of agency and self-ownership is incomplete with respect to the avatar. The information produced and transmitted inside the VE cannot reach all the senses. All the avatars can be perceived as negotiable bodies, but most unforgettable experiences are reserved for the immersive VR environment. The linguistic intuitions of users also seem to support the hypothesis that the virtual self, namely one's avatar, is regarded as an extension of the user. The use of the first-person pronoun in these kinds of contexts is common enough (Velleman 2013, 14). When narrating their virtual encounters and experiences, users refer to their avatar with the first-person pronoun, saying 'I did this or that', instead of 'My avatar did this or that'.

The peculiar features of 'I' generated a fertile debate between philosophers in what concerns its ability to successfully self-refer. Wittgenstein distinguished between two uses for the first-person pronoun: as a subject, when the speaker intends to communicate information about her beliefs and emotions, for example, and as an object, when the speaker tries "to match up first-person experience with some known criterion in order to judge the experience to be [her] own" (Gallagher 2000, 15). When a player says 'I like video games', it would be nonsensical to question the ability of the speaker to correctly identify herself. The subjective use of the first-person pronoun is immune to error through misidentification (Shoemaker 1968; Evans 1982).

On the other hand, when a VR user says 'I am on top of Mount Everest', she can fail to refer successfully to her own person. Let us suppose that someone will spend the entire morning in her room logged in a VR application, exploring Mount Everest and trying to overcome her fear of heights. She will say 'I was on Mount Everest', but she would objectively misidentify herself. She was in her room, immersed in a VR game, while her avatar was on Mount Everest. We can accept that her body was not located on Mount Everest at the indicated time, and therefore the sentence is false. But do we have sufficient reason

[1] The Rubber Hand Illusion and the Virtual Hand Illusion are induced by applying repeated and synchronized strokes to a rubber hand/ virtual hand that is positioned so that it appears to be an extension of the bodily self of the participant in the experiment.

to say that the first-person pronoun in used erroneous in this case? The answer to this question is dependent on the weight one is willing to give to occurrent moral and linguistic intuitions. It is true that the VR user is not physically on Mount Everest, but all of her senses were tricked into believing that she is there, experiencing avatar embodiment, acting according to her beliefs and desires. From this perspective, the objective use of the first-person pronoun can be taken to encompass more than the mere location of the body. Through immersion, the user experiences the virtual environment as a substitute for reality.

The subjective use of the first-person pronoun is less philosophically problematic. When the VR user says ‘I was very afraid I would fall’, the use of the first-person pronoun is immune to misidentification. The privileged access one has to its own first-person experiences and to the phenomenological content of those experiences cannot be incorrectly attributed by their possessor. Even if the VR user’s body is safe inside the confinement of her room, the user is present inside the virtual environment generated by a VR equipment.

To better understand why VR users tend to use the first-person pronoun when referring to their avatars, we must turn our attention to avatars and their standing in relation to ourselves.

III. Avatars as negotiable bodies

The virtual environment is a digital space populated by diverse virtual objects, including human and animal representations. Human representations are either called avatars, when controlled by a human user, or agents, when controlled by an algorithm (Fox et al. 2009, 97). The object of interest for this present paper is the avatar, the “pictorial constructs used to actually inhabit the [virtual] world” (Taylor 2002, 40). Avatars can be highly realistic, pictorials designed to recreate the exact physical appearance of a user through facial modelling techniques. At the other end of the spectrum, one can manipulate a pixelated avatar that looks nothing like the user controlling and incorporating it in the virtual environment. The degree of realism imprinted on an avatar has little to no bearing on the degree of immersion and presence the user is experiencing in VR settings. With the exception of the games designed to randomly assign an avatar to the player, the avatar is an expression of one’s form of choice for the virtual embodiment.

Choosing and customizing an avatar can be seen as a technique of negotiating the boundaries of the self. When talking about the use of prosthetic devices in the same way in which one makes use of natural limbs, Andy Clark notes that “creatures capable of this kind of deep incorporation of new bodily structure are examples of what I shall call “profoundly embodied agents”. Such agents are able constantly to negotiate and renegotiate the agent-world boundary itself” (2008, 34). The avatar is commanded just like a prosthetic arm, an alien addition to the natural body, who is manipulated through automatic, unreflective brain commands. When she/he manipulates a prosthetic arm to hold a grab a book, she/he does not say “My prosthetic arm grabbed a book”, but rather “I grabbed a book”. The same phenomena appear in the case of avatars- the users typically say ‘I did this or that” because the mind is immersed in the virtual world when the act occurs. The avatar does not appear in the practical reasoning of VR users because the coordination of brain command and movements is accurate enough.[2] The development of VR equipment designed to enhance the immersion of users into the virtual environment, such as haptic gloves and full-body haptic suits, makes possible the complete coordination between the virtual body of the avatar and the physical body of the user.

In this sense, the avatar as a negotiated body provides "access points in the creation of identity and social life. The bodies people use in these spaces provide a means to live digitally- to fully inhabit the world. It is not simply that users exist as just 'mind', but instead construct their identities through avatars" (Taylor 2002, 40). For Taylor, the avatar is a fictional self who can act as a means of self-discovery and self-creation for players. The Virtual Reality is not solely experienced by users' disembodied minds, roaming freely in a digitally rendered space, but by embodied agents, virtual extensions of the self. The non-conceptual first-person content displayed by users in VR also indicates to individuals assuming an embodied position in the environment. This type of content manifests like the feeling of being there, in the virtual environment, and can be seen as a manifestation of self-presence. The reports of Second Life players are testimony to an even more robust sense of selfhood experienced through one's avatar. Some SL residents believe that virtual embodiment is a way of expressing the authentic self, saying that the avatar reflects the way in which they perceive themselves to be "on the inside" (Boellstorff 2008, 134). For example, those residents who are body bound by permanent physical disabilities are negotiating the form of their embodiment (137), thus exercising ways of interacting with the environment that are not available to them in the physical world. Nonetheless, the subjective experience one has when manipulating an avatar in Second Life is not bound to be coherent. The practice of having and using alts (more than one avatar per player in the same virtual environment) simultaneously can cause fractures in users' subjectivity (150). The same objection cannot be made in respect to VR experiences because VR users cannot virtually embody more than one avatar at a time.

Virtual embodiment is perceived as being an authentic form of expressing one's self especially in the presence of virtual communities. A virtual world design that favors inter-subjectivity creates the conditions of possibility for the users to experience a full social life in a virtual space. In the absence of inter-subjectivity, a user can only experience the ecological space rendered by the hardware setup. Even though "the plasticity of our self-representations" and the characteristics of the virtual environment are very important to the creation of online identities, "technical affordances on social interaction in online environments" can reduce the gap between real and virtual selves (Yee and Bailenson 2007, 272).

IV. From the extended mind to the extended self

In their highly influential article entitled "The extended mind", Andy Clark and David Chalmers argued for an active form of externalism about the mind, a view which entailed that cognitive processes supersede the traditionally accepted boundaries of the skin and

[2] The argument is inspired by David Velleman's treatment of players controlling avatars through the keyboard and mouse. Velleman believes that, as the player gains more skills in controlling her avatar, the manipulation of artifacts disappears from her explicit intentions (2013: 12).

skull (1998, 7). If, when dealing with epistemic actions, an artifact functions as a process which we would normally recognize as being a cognitive process if it were to happen inside the head, then the artifact-mediated process is also a part of the cognitive process (8). To illustrate the extended mind thesis, Clark and Chalmers use the example of Otto, an Alzheimer's patient, who extends his mind into the world through the means of his notebook. The notebook acts as Otto's surrogate memory: every new piece of information or belief about the world is written down and looked up in the notebook by Otto when needed (12). The notebook here has the same function as one's biological memory.

The extended mind thesis, in its original form, consists of three argumentative layers: one arguing in favor of extending cognitive processes into the world, one regarding the extension of cognitive states, and the last one concerning the self. The existence of an extended mind carries with it an extension of the concept of person as an "integrated system when coupled with external resources" (Shin 2013, 83). If the mind is not bound by the limits of the skull, personal identity can also be seen as an extended system, "spread into the world" (Clark and Chalmers 1998, 18). The rapid pace of innovation in technology may call for this sort of reconceptualization of personal identity and the boundaries of the self. The use of prosthetic devices meant to enhance the human body or to remedy a lack in the ability to perform certain actions, ranging from pragmatic to epistemic actions, is already a matter of great interest for philosophers. But one of the newest puzzles for personal identity theoreticians is the role that our virtual lives have in who we are and how we understand ourselves. The extended mind thesis invites to a reinterpretation of the relation between virtual embodiment and the creation of the self in rapidly evolving world of immersive technologies. Building on the model put forward by Clark and Chalmers, the extended theory of identity should take the following form: *if we would normally see a particular action or process as being a part of one's narrative identity when performed by a physical body in the physical world, then the corresponding process performed by the corresponding virtual body in a virtual environment, controlled by the same mind, must receive the same philosophical treatment*. In other words, one's avatar must be regarded as the extended self of the user.

With this in mind, we can alter Otto's example in order to better understand the mechanism of mind and self extension into the world, be it virtual or physical, and the ethical implications of adhering to such a position.

Let us imagine Otto logging into Sansar from his room. Once he has the VR headset on, his haptic gloves and the full-body haptic suit, he is fully immersed in the virtual environment. Little to no external stimuli would be registered by Otto's senses. He built for himself an avatar called Toto to resemble his physical appearance. Otto knows that Toto's adventures can cause him no real physical harm: even if Toto can get hurt and even die in his virtual universe, Otto would still be safe in his room. If Otto would have a heart attack and die, his avatar would not. It appears that Otto and Toto are two distinct entities. But Toto is mimicking every move Otto does in the real world. If Otto moves his hand to the right, Toto would do the same. If Toto is hit by an object in Sansar, then Otto would feel his full-body haptic suit vibrating. It is the case that" the participant in a virtual world moves his avatar under the impetus of his own beliefs and desires about the virtual world, and he does so with intentions like the ones with which he moves his own body (and its prosthetic extensions) under the impetus of his beliefs and desires "(Velleman 2013, 15). Otto's narratives- the real life narrative and the virtual odyssey associated with his avatar- intertwine inasmuch they are "subplots in the more comprehensive narrative of the resident (...) Both sets of adventures are part of the same life because, although distinguishable sub-narratives, they impact each other along the most fundamental dimensions of narrative interaction" (Schechtman 2012, 341). Interpreting this phenomenon through a narrativist take on personal identity, we arrive at the conclusion that Otto cannot insulate his real-life narrative from that of his avatar. In turn, the latest can have transformative effects on the meta-narrative in the same degree as the real-life narrative.

At this point, a critic could formulate the following objection: a lot of external factors impact one's process of constructing the narrative self, and these factors cannot be regarded as extensions of selves. Books, fictional characters that we inhabit in our imagination while reading a story, can be taken to have a similar impact on the construction of one's character. Whilst fictional characters can play a fundamental part in character formation, the relationship between a reader and a character is very different from that between a VR user and her avatar. As readers, we can be passive or engaged recipients of experiences that make an impression on us because one can exercise sympathy towards the characters and their narrative arc. We exercise our moral capacities only to pass judgements on character's choices and behavior, to enjoy or dislike the course of action, etc. A reader is never an agent in a story because the narrative is controlling the experience, not the one who reads it. VR users are in complete control of the narrative. They have absolute control over their actions and here resides the strongest, yet trivial, answer to such an objection.

I have argued, up to this point, that VR experiences and avatars have a special status when compared to other potentially character-building environments. This argumentative step is necessary, but not sufficient in supporting an extended identity view. In order to be successful in showing how one's self can reside inside the avatar, I argue that, more often than not, these experiences have a long-lasting, real impact on the way users perceive and understand themselves.

Whilst the physical harm cannot transcend the immersive virtual reality setup, other type of experience outcomes might spill over in the real world. Otto can feel happiness, sadness, excitement or anxiety as a consequence of his virtual activities and interactions because he is the one to live them- Toto is a mere depository of Otto's agency. In comparison with the avatars from computer games, VR avatars are stimulating a sense of agency and ownership in the human users behind them.

The possibility of living unforgettable experiences in VR is conditioned by the induced belief that users are the ones causing an action, not the avatar. In fact, the experience could not belong to anybody else besides myself, the user, as the avatar has no memory, intentionality and no standing agency.

What counts as an unforgettable life experience? It is surely not just the actions done under the fear of suffering physical pain. Jacquelyn F. Morie takes an unforgettable experience to be the kind that "lasts beyond the time of the actual experience. It could initiate the formation of strong memories of the experience, reignite ties to personal memories, or initiate a lasting change within the experiencer" (2006, 2). In other words, for something to count as a meaningful experience, it must be able to substantiate a psychological effect in the experiencer that supersizes the conceptual limits of a minimal self.[3]

[3] The minimal self refers to "the consciousness of oneself as an immediate subject of experience, unextended in time" (Gallagher 2000: 15).

One particular aspect of the virtual social life can be of use in trying to illustrate the heavy weight of these experiences in relation to one's autobiographical sense of self- experiencing virtual intimacy through an avatar. Many residents of SL report engaging in relationships with other avatars. When things get sour, "the sense of loss could be as intense as with an actual-world relationship" (Boellstorff 2008, 173). Players also engage frequently in sexual activities, ranging from sex work to public orgies (161). Some of these sexual experiences have a strong linguistic component. In VR, the movements of one's body give content to an interaction. The user has complete and immediate control over her actions. The rise of new technologies will transform the way in which we explore our sexuality, allowing more complex interactions between avatars and the corresponding bodies from the physical world. One report predicts that, in the following years, new devices will transform sexual virtual interactions into experiences more real than ever: the use of integrated sex devices, immersive VR with touch, cybersex, connected dildos and sex sleeves, long-distance kissing devices, etc. (Owsianik and Dawson 2017). This, in turn, will affect the quality of the sexual experiences in VR worlds, which would better emulate the real experience. VR will then become an even more accessible space for self-discovery, where the boundaries of the virtual world will be bent by the use of devices and brain interfaces.

VR applications already are genuine spaces for self-discovery, cognitive and moral enhancement. Some experimental applications of VR show that, when a subject is induced the illusion of full body ownership with respect to her virtual body, the socio-perceptual processes of users can be substantially modified. One of these experiments shows that VR technology can be effectively used to help domestic violence offenders learn to identify emotional responses and practice their moral capacity of sympathy towards victims of abuse (Seinfeld et al. 2018). Such a result was possible due to the shift in perspective- the offenders virtually embodied female avatars who were subjected to abusive treatments. If one would have a similar experience in the real life, that said experience would be seen, without a doubt, as being fundamental to character development, a non-invasive technique of moral enhancement or moral treatment.

V. Further remarks

In this paper, I explored the possibility of expanding our notion of selfhood into immersive virtual worlds that use first-person perspective, rather than a view from one's avatar proximity. Even if I did not support the thesis that VR environments are inherently different from other virtual environments, I took the first-person perspective and the technological ability of coordinating the body movements of users and avatars (matching the proprioceptive feedback through body-motion capture) as being fundamental for reaching a state of hyper-immersion (Miller and Bugnariu 2016), which in turn facilitates the identification of one's self with the avatar (self-presence). Virtual experiences can have a lasting psychological effect on the users. The reports of Second Life residents show that the relations they form in the virtual space and the design of the platform, which fosters meaningful social interactions, provide the premises for self-reported authentic life experiences. VR offers an even more immersive space one can inhabit, a genuine reality engine in which users can experience love, friendship, sex and intimacy.

IVR technology holds great promise for non-invasive cognitive and moral enhancement, whilst also exposing users to virtual harms. It operates in a "reality horizon" (Slater and Sanchez-Vives 2016) and, like any other technology, it is both a burden and a blessing (Postman 1993, 5). An extended selfhood thesis bears wide ethical implications that must be further explored. If we accept that the avatars one inhabits in IVR social worlds are part of who we are, we must also consider how one can evaluate the moral responsibility one bears for virtual moral transgressions that are already happening, such as murder, rape (Marika 2019), theft, and other forms of physical and emotional harm performed in the virtual reality.

We externalize more and more cognitive processes into the external world through various devices, starting with our memory. It is not implausible to think that, in the near future, we would rely more and more on virtual environments for self-discovery and experimentation of all sorts. This calls for conceptual refinement of notions such as personal identity, moral responsibility and social life. This paper was an attempt to do just that.

References

Biocca, Frank, and Ben Delaney. "Immersive virtual reality technology". In *Communication in the age of virtual reality*, edited by Frank Biocca and Mark Levy, L. Erlbaum Associates Inc.365 Broadway Hillsdale:NJ, 1995.

Biocca, Frank, and Mark R. Levy. "*Communication in the age of virtual reality*". L. Erlbaum Associates Inc.365 Broadway Hillsdale:NJ, 1995.

Boellstorff, Tom. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. New Jersey: Princeton University Press, 2008.

Borrego, Adrian, Jorge Latorre, Mariano Alcaniz and Roberto Llorens. "Embodiment and Presence in Virtual Reality after Stroke. A Comparative Study with Healthy Subjects", *Frontiers in Neurology* 10(2019). https://doi.org/10.3389/fneur.2019.01061.

Buller, Tom. "Neurotechnology, Invasiveness and the Extended Mind." *Neuroethics* 6, no.3 (2013): 593-605. https://doi.org/10.1007/s12152-011-9133-5.

Carlson, Matthew, and Logan Taylor. "Me and My Avatar: Player-Character as Fictional Proxy", *Journal of the Philosophy of Games* 2, no. 1(2019):1-19. https://doi.org/10.5617/jpg.6230.

Clark, Andy, and David Chalmers. "The extended mind". *Analysis* 58, no.1 (1998): 7-19.

Clark, Andy. "*Supersizing the Mind*". New York: Oxford University Press, 2008.

Evans, Gareth. "*The Varieties of Reference*". New York: Oxford University Press, 1982.

Fox, Jesse, Dylan Arena, and Jeremy Bailenson. "Virtual Reality: A survival guide for the social scientist". *Journal of Media Psychology*, 21 (3), 2009:95-113. 10.1027/1864-1105.21.3.95.

Gallagher, Shaun. "Philosophical Conceptions of the Self: Implications for Cognitive Science." *Trends in Cognitive Sciences* 4, no. 1(2000): 14-21. 10.1016/s1364-6613(99)01417-5.

Gonzalez-Franco, Mar, Tabitha Peck, Antoni Rodriguez-Fornells, and Mel Slater. "A threat to a virtual hand elicits motor cortex activation". *Experimental brain research* 232, no. 3 (2013):875-887. 10.1007/s00221-013-3800-1.

Gotlib, Anna. (2014). "Girl, Pixelated- Narrative Identity, Virtual Embodiment, and Second Life." *Humana.Mente: Journal of Philosophical Studies* 7, no. 26 (2014): 152-178. Retrieved from http://www.humanamente.eu/index.php/HM/article/view/120.

Lee, Kwan Min. "Presence, Explicated", *Communication Theory* 14, no. 1 (2004): 27–50. https://doi.org/10.1111/j.1468-2885.2004.tb00302.x.

Macedonio, Mary, Thomas Parsons, Raymond Digiuseppe, Brenda Weiderhold, and Albert Rizzo. "Immersiveness and physiological arousal within panoramic video-based virtual reality". *CyberPsychology and Behavior*, 10, 2007: 508–515. 10.1089/cpb.2007.9997.

Madary, Michael, and Thomas Metzinger. "Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology". *Frontiers in Robotics and AI*, 3 (2016). https://doi.org/10.3389/frobt.2016.00003.

Marika, Guggisberg. "'Rape day'- A virtual Reality Video Game Causes Outrage". *Psychol Psychother Res Stud* 2, no. 3 (2019). 10.31031/PPRS.2019.02.000537.

Metzinger, Thomas. "*Being No One: the self-model theory of subjectivity*". MIT Press, Massachusetts, United States, 2003.

Miller, L. Haylie and Nicoleta L. Bugnariu. "Levels of Immersion in Virtual Environments Impacts the Ability to Assess and Teach Social Skills in Autism Spectrum Disorder". *Cyberpsychol Behav Soc Netw*. 19, no.4 (2016): 246–256. 10.1089/cyber.2014.0682.

Morie, Jacquelyn Ford. "Virtual reality, immersion and the unforgettable experience". *SPIE-IS&T Proceedings* 6055(2006). 10.1117/12.660290.

Popat, Taylor Sita, and Kelly Preece. "Pluralistic presence: practicing embodiment with my avatar". In *Identity, Performance and Technology: Practices of Empowerment, Embodiment and Technicity*, edited by Susan Broadhurst and Josephine Machon, Hampshire, Great Britain: Palgrave-Macmillan, 2012. 10.1057/9781137284440_11.

Postman, Neil. "The Judgement of Thamus". In *Technopoly: The Surrender of Culture to Technology*, 3-21. New York:Vintage Books, Random House Inc., 1993.

Riva, Giuseppe, Fabrizia Mantovani, Claret Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz Raya. "Affective interactions using virtual reality: The link between presence and emotions". *CyberPsychology and Behavior*, 10, no.1 (2017):45–56. 1089/cpb.2006.9993.

Owsianik, Jenna, and Ross Dawson. "The Future of Sex Report", 2016. https://futureofsex.net/Future_of_Sex_Report.pdf.

Robson, Jon, and Aaron Meskin. "Video Games as Self-involving Interactive Fictions". *The Journal of Aesthetics and Art Criticism* 74, no.2(2016):165-177. https://doi.org/10.1111/jaac.12269.

Schechtman, Marya. "The Story of my (Second) Life: Virtual Worlds and Narrative Identity." *Philosophy & Technology*, 25 (2012): 329–343. 10.1007/s13347-012-0062-y

Seinfeld, Sofia, Jorje Arroyo-Palacios, G. Iruretagoyena et al. "Offenders become the victim in virtual reality: impact of changing perspective in domestic violence." *Scientific Reports* 8(2018): 2692. https://doi.org/10.1038/s41598-018-19987-7.

Shin, Sangkyu. "Extended Mind and the Extension of a Self." *Lyceum* 1(2013): 81-100.

Shoemaker, Sydney. *Identity, Cause, and Mind*, New York: Oxford University Press, 2003 [1984].

Slater, Mel, Marcos Daniel Pérez, Henrik Ehrsson, and Maria V. Sanchez-Vives. "Inducing illusory ownership of a virtual body". *Frontiers in Neuroscience*. 3(2009),2:214- 220. 10.3389/neuro.01.029.2009.

Slater, Mel and Maria Sanchez-Vives. "Enhancing Our Lives with Immersive Virtual Reality". *Frontiers in Robotics and AI*, 3(2016):1-74. https://doi.org/10.3389/frobt.2016.00074.

Spanlang, Bernhard, Normand, Jean-Marie, Borland, David, Kilteni, Konstantina, Giannopoulos, Elias, Ausiàs Pomés, Mar Gonzalez-Franco, Daniel Perez-Marcos, Jorge Arroyo Palacios, Xavi Navarro, and Mel Slater. "How to Build an Embodiment Lab: Achieving Body Representation Illusions in Virtual Reality." *Frontiers in Robotics and AI*, 1(2014): 1-22. https://doi.org/10.3389/frobt.2014.00009.

Takahashi, Dean. "Sansar gets revamped with avatar editor, a new core world, and corporate partnership", *Venture Beat*, September 24, 2019. https://venturebeat.com/2019/09/24/sansar-reimagined-with-avatar-editor-a-new-core-world-and-corporate-partnerships/.

Taylor, T. L. (2002). “Living Digitally: Embodiment in Virtual Worlds.” In *The Social Life of Avatars: Presence and interaction in shaded virtual environments*, edited by Ralph Schroeder, 40-62. London: Springer, 2002.

Velleman, J.David. “Virtual Selves”. In *Foundations for Moral Relativism*, 5-23. Cambridge: Open Book Publishers, 2013.

Yee, Nick and Jeremy Bailenson. “The Proteus Effect: The Effect of Transformed Self Representation on Behavior”. *Human Communication Research*. 33(2007):271-290. 10.1111/j.1468-2958.2007.00299.x.

Yuan, Yee and Anthony Steed. “Is the rubber hand illusion induced by immersive virtual reality?” In *2010 IEEE Virtual Reality Conference (VR)*, Waltham, MA (2010): 95-102.

Author information
Anda Zahiu
Research Center in Applied Ethics, University of Bucharest