Scientia Moralitas
International Journal of Multidisciplinary Research
Vol. 4, No. 1, 2019

Editors: Dr. Ioan–Gheorghe Rotaru
Dr. Julia M. Puaschunder


SCIENTIA


MORALITAS

SCIENTIA MORALITAS
Vol. 4, No. 1 | Year 2019

Scientia Moralitas Research Institute
Beltsville, MD, 20705, USA
www.scientiamoralitas.org
E-mail: scientia.moralitas@email.com

Ioan–Gheorghe **Rotaru**, Editor-in-Chief
Julia M. Puaschunder, Guest Editor

LIBRARY OF CONGRESS * U.S. ISSN CENTER
ISSN 2472-5331 (Print)
ISSN 2472-5358 (Online)

First published, 2016

SM SCIENTIA MORALITAS

CONTENTS

EDITORIAL BOARD

Scientia Moralitas International Journal of Multidisciplinary Research
ISSN 2472-5331 (Print) | ISSN 2472-5358 (Online) | Vol. 4, No. 1, 2019
DOI:

A Behavioral Approach to Irrational Exuberances – An Artificial Intelligence Roboethics Taxonomy

Dirk Beerbaum, PhD
Aalto University School of Business,
Department of Accounting, Helsinki, Finland, Dirk.Beerbaum@aalto.fi
Frankfurt School of Finance & Management,
Frankfurt am Main, dbeerbaum@fs.de

Julia M. Puaschunder, PhD
The New School, Department of Economics,
Schwartz Center for Economic Policy Analysis,
Columbia University, Graduate School of Arts, New York,
Julia.Puaschunder@columbia.edu,
Princeton University, Julia.Puaschunder@princeton.edu,
George Washington University, jpuaschunder@gwu.edu

> "Clearly, sustained low inflation implies less uncertainty about the future, and lower risk premiums imply higher prices of stocks and other earning assets. We can see that in the inverse relationship exhibited by price/earnings ratios and the rate of inflation in the past. But how do we know when irrational exuberance has unduly escalated asset values, which then become subject to unexpected and prolonged contractions as they have in Japan over the past decade?" (Alan Greenspan, 1996)

ABSTRACT: Contemporary theories and studies of economics apply a behavioral approach. Behavioral Economics revolutionized mainstream neo-classical economics in the past years. The success of behavioral economics is reflected by two Nobel Prizes in Economics. The wide range of psychological, economic and sociological laboratory and field experiments proved human beings

deviating from rational choices and standard neo-classical profit maximization axioms often failed to explain how human actual behavior. Human beings rather use heuristics in their day-to-day decision making. These mental short cuts enable to cope with a complex world yet also often leave individuals biased and falling astray to decision making failures. Artificial intelligence (AI) driven robots and machines are forecasted to grow dramatically in the next years . AI reflects many algorithms, models and techniques, machine learning, databases and visualizations. One of the main advantages of AI-driven machines is that they follow consistently rational algorithmic rules without being biased. Ethical considerations intend to make AI-driven robots more human and introduce morality into machines. The Uber-Waymo trial made transparent how much artificial intelligence development is impacted by human irrationality and irrational exuberances. It reveals a culture of agile software development, which prioritize releasing the latest software over testing and verification, and one that encourages shortcuts and irrationality. This also give proof that applying artificial intelligence cannot ensure that irrational exuberances are prevented. The reason for this irrational exuberance may have its roots in the exponential growth in computing and storage technologies predicted by Gordon Moore five decades ago. This paper develops a concept how irrational exuberances can be prevented from happening. One general approach for solutioning of the issue is to increase transparency. The paper recommends applying technology to make data more accessible and more readable on the application of artificial intelligence. For this purpose the application of"transparency technology XBRL (eXtensible Business Reporting Language)" is incorporated. XBRL is part of the choice architecture on regulation by governments (Sunstein 2013), which applies nudging for influencing towards a preferred option used by the mass consumers. XBRL is connected to a taxonomy. The paper develops a taxonomy to make application of artificial intelligence more transparent to the public and incorporates ethical considerations. As a business case the strongly growing robo-advice market in Germany is taken. The taxonomy is either inductively derived from the robo-advice market offerings and deductively includes the existing standards on ethical codes for robot's usage and application of artificial intelligence. The paper focus on the way to enhance AI that aligns with human values. How can incentive be provided that AI systems themselves do not become potential objects of moral concern. The main outcome of the paper is that Digitalization implies with AI moral concerns however transparency technologies at the same time also offer way to mitigate such risks.

KEY WORDS: Irrational exuberances, Artificial Intelligence Ethics, Behavioural Economics, Human-Computer Interaction, Taxonomy, XBRL and Transparency

Introduction

Contemporary theories and studies of economics apply a behavioral research approach. This is underpinned by the fact that behavioral economics reversed mainstream neo-classical economics in 21st century. Since then two Nobel Prizes in Economics were distributed as a wide range of psychological, economic and sociological laboratory and field experiments proved human beings deviating from rational choices and standard neo-classical profit maximization axioms often do not constitute explanations for human behavior. Human instead of pure rationality rather apply heuristics in their day-to-day decision making. These mental deficiencies often leave individuals incapable of avoiding decision making failures within a complex world. Research e.g. in Political Science about voting decision from people give proof that people are strongly influenced by rather unreflective first impressions and as a result decisions based on that are not driven by rational reflections and deliberations .

Behavioral Economics intend to specify anomalies and shortfalls in neo-classical economics. Due to mental deficiencies, humans are incapable to guide their lives proactively within a complex world and rather become victim and tributary to complexity. Opposite to the assumptions of the standard neo-classical theory, individuals intend to reduce complexity, whenever the opportunity is provided , which reflect irrational exuberances. Irrational exuberances are well described in Shiller's book about the housing market "The market is high because of the combined effect of a lot of indifferent thinking across millions of people, very few of whom feel a need to do careful research about the long-term investment value of the aggregate stock market, and who are motivated substantially by their own emotions, random attentions, and perceptions of conventional wisdom. Their behavior is heavily influenced by news media that are interested in attracting viewers or readers, with little incentive to report regularly on quantitative analysis that might give a correct impression of the aggregate stock market level." Reducing complexity also implies decreasing cognitive drain on mental resources. For many day-to-day problems, humans develop certain heuristics as in Shiller's description on the appreciation of the housing market, which represent mental simplifications or rule of thumbs . Contrary to neo-classical

assumptions, pareto optimality for society over time does not become in conformity with the aggregated individual generations' preferences, as the sum of individual generations' preferences will not lead to societally favorable outcomes over time .

Due to this conflict, behavioral economists have recently started to nudge – and most recently wink – people into favorable decision outcomes, offering promising avenues to steer social responsibility in public affairs. The freedom of economic choice and the assumption that free markets lead to efficient outcomes, which is often described in the literature with the metaphor of Adam Smith invisible hand is questioned due to human irrationality. This new idea of interfering into the market became very sucessful and was extended to different fields. What followed was the powerful extension of behavioral insights for public policy making, international development and decision usefulness. Behavioral economists proposed to nudge and wink citizens to make better choices for them and the community around the globe. Many different applications of rational coordination followed ranging from improved organ donations, health, wealth and time management, to name a few. Starting with the beginning of the entrance of behavioral aspects in economic analyses and intercultural differences in behavioral understandings, the paper will then embark on a wide range of classic behavioral economics extensions in order to guide a powerful application to AI in the age of the digitalization of the economy.

This paper applies behavioral economics to an issue appearing in the area of investor decision usefulness caused by the digitalization of the economy in a truly interdisciplinary way. What role do ethics play for behavioral economists? In the future age of AI, should we create algorithms that resemble human decision making or strive for rational artificiality? Can transparency technology such as XBRL help to counteract against the associated risk of unethical application of AI? And does nudging in the wake of libertarian paternalism entail a social class division into those who nudge and those who are nudged? This paper develops based on AI-driven products in the Banking and Finance Industry such as Roboadvisors and AI-driven finance robots, a taxonomy that reflects ethical consideration and upon application enables a way to mitigate such risks by providing enhanced transparency.

Artificial Intelligence (AI)

Artificial Intelligence (AI) implies historically unique opportunities but also threats to humankind. As an emerging global trend, AI becomes relevant at almost all levels of social conduct and thereby raised both – high expectations but also grave concerns. AI reflects many algorithms, models and techniques, machine learning, databases and visualizations. One of the main advantages of AI-driven machines is that they follow consistently rational algorithmic rules without being biased. Ethical considerations intend to make AI-driven robots more human and introduce morality into machines. The Uber-Waymo trial made transparent how much artificial intelligence development is impacted by human irrationality and irrational exuberances.

This also give proof that applying AI cannot ensure that irrational exuberances are prevented. The reason for this irrational exuberance may have its roots in the exponential growth in computing and storage technologies predicted by Gordon Moore five decades ago. With the dramatic increase in diversity and the usage of emerging technologies in today's societies, such as social robots, lifelike computer graphics (avatars), virtual reality tools and haptic systems and Roboadvisors the social complexity of these challenges are rising . One of the main challenges in developing and applying modern technologies in our societies is the treatment of ethical issues surrounding AI (Meghdari and Alemi 2018). The call for AI Ethics (AIE) has emerged e.g. reflected by the European Group on Ethics in Science and New Technologies. It reveals a culture of agile software development, which prioritize releasing the latest software over testing and verification, and one that encourages shortcuts and irrationality.

A growing number of AI and robotics researchers have expressed their willingness and the requirement to create a framework on AI ethics building on the benefits of humanities, philosophy, natural sciences, sociology, and social neuroscience. AI enables the potential to replicate human existence but with indefinite lifetime. From the view of overpopulation concerns, under the assumption that AI can help to substitute machines for humans AI would be a solution to avoid a crowding of the planet. AI currently also reaches quasi-human status through actual personhood – e.g., via citizenship

and quasi-human rights applied in the Common Law but also Roman Law territories of the US and the EU. Leveraging AI entities to the status of being through the attribution of legal personhood raises challenging legal and ethical questions. A novel predicament between eternity and overpopulation hence calls for revising legal codes for killing and ethical imperatives and religious concerns over suicide.

AI consist of a large number of algorithms, models and techniques, machine learning, databases and visualizations . According to AI is the science and engineering of producing intelligent machines, particularly computer programs, which incorporate intelligence and implies also the task of using computers to understand human intelligence. Historically, the process leading to the enormous spread of information and technology is frequently considered as the digital revolution. The term reflects a revolutionary development from the industrial to the information age. This transition towards economies and business models implies the usage of information and communication technology and virtual processes instead of analogue mechanics and face-to-face services (Moudud-Ul-Huq 2014). The second half of the last century was dominated by the development of computer technology. This is often referred to as the Third Industrial Revolution, which was driven by the invention of microprocessors that enabled the mass production of personal computers and a very fast increase in storage and computing capacity . As the most novel trend, AI, robots and algorithms are believed to soon disrupt the economy and employment patterns. With the advancement of technologies, employment patterns will shift to a polarization between AI's rationality and humanness. Robots and social machines have already replaced people in a variety of jobs – e.g. airports smart flight check-in kiosks or self-check-outs instead of traditional cashiers. Almost all traditional professionals are prospected to be infused with or influenced by AI, algorithms and robotics. For instance, robots have already begun to serve in the medical and health care profession, law and–of course–IT, transportation, retail, logistics and finance, to name a few. Social robotics may also serve as quasi-servants that overwhelmingly impact our relationships.

AI's entrance in society will revolutionize the interaction between humans and AI with amply legal, moral and social implications . Autonomous

AI entities are currently on the way to become as legal quasi-human beings, hence self-rule autonomous entities . AI can in principle be distinguished between weak AI, where "the computer is merely an instrument for investigating cognitive processes" and strong AI, where "[t]he processes in the computer are intellectual, self-learning processes". Weak AI is labeled as Artificial Narrow Intelligence (ANI) while strong AI is further distinguished between Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI).

Exponential growth in data availability enabled the development of AI systems for pattern selection in big data and a broad range of applications, such as speech and natural language processing, computer vision, image recognition (e.g. in search engines and social networks) and predictive analytics. This founded the basis for virtual personal assistants such as Alexa, Siri or Cortana, which have become first AI-enabled tools used by the mass consumers. Remarkable is the speed with which these radical changes are occurring, and their extensive and comprehensive systemic proliferation have become known as the Fourth Industrial Revolution, as popularized by World Economic Forum founder Klaus Schwab. The pace of technological development has gained such speed that corporates, consumers and governments often find themselves struggling to keep pace. Developments in AI have far-reaching economic and sociopolitical consequences, some of them are already materializing (Körner 2018). However, it is still unclear, what will be the exact impact on human society. How will AI and robotics lead to the allocation of labor and capital? When people decide, limitations in their capacity to foresee long-term impacts and the collective outcomes of their choices can contribute to institutional downfalls. The more machine learning systems apply AI becomes powerful it will become more important that ethical frameworks are incorporated. According to machine learning are computational algorithms that use certain characteristics to learn from data using a model.

It has been long history since society was concerned with the impact of robotics technology. From nearly a century ago the word "Robot" was mentioned for the first time . The EU Committee on Legal Affairs (2016, 4) holds that "[U]ltimately there is a possibility that within the space of

a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also to its capacity to be in charge of its own destiny and to ensure the survival of the species." AI mimicking human intellect could soon surpass humans intellectually but also holistically breaking the barrier of human controlled-automization (Schuller 2017). Modern literature about robots features cautionary accounts about insufficient programming, evolving behavior, errors, and other issues that make robots unpredictable and potentially risky or dangerous. "Observe, orient, decide, act" will therefore become essential in the eye of machine learning autonomy and AI forming a new domain of intellectual entities (Armstrong & Sotala 2012, 52; Copeland 2000; Galeon & Reedy 2017; Marra & McNeil 2013). The uncertainty surrounding AI development and self-learning capabilities give rise to the need for guarding AI and an extension of the current legal system to cope with AI (Themistoklis 2018).

With the advancement of technology, social robots have found broader applications in the private and public sectors, such as educational and cultural affairs, games and entertainment, clinical and rehabilitation, nursing of children and/or elderly, search and rescue operations). For example, social robots such as ASIMO, Nao, iCub, ARASH, and RASA have been developed for "Edutainment" or "education entertainment" purposes. They aid the study of cognition (both human and artificial), motion, and other areas related to the advancement of robotics serving our society (Meghdari and Alemi 2018). In addition, a few medical and healthcare toy-like robots, such as PARO, which looks like a baby seal, or ARASH, which is a humanoid, have been designed for therapeutic purposes such as reducing distress, stimulating cognitive activity, teaching specific subjects, and improving socialization (Meghdari and Alemi 2018). Similarly, Sharif University of Technology's socially assistive robot RASA has been developed to help coach and teach Persian Sign-Language to Iranian deaf children (Meghdari and Alemi 2018). Personal care and companion robots are increasingly being used to care for the elderly and children, such as RI-MAN, PaPeRo, and CareBot (Meghdari and Alemi 2018). In recent years, robotics technology has extended its applications from factories to more general-purpose practices in

society – for instance, such as the use of robots in clinical and rehabilitation, nursing and elderly care, search and rescue operations (Meghdari and Alemi 2018). Social robots have become clinical and educational assistants for social interventions, treatment, and education such as language trainings but also assistance with children with disabilities like autism, down syndrome, cancer distress, hearing impairment, etc. (Meghdari and Alemi 2018). Initial investigations clearly indicate that social robots can play a positive role in the improvement of children's social performance, reduction of distress during treatments, and enhancing their learning abilities (Meghdari and Alemi 2018). Surprisingly, although not too hard to imagine, relationships of a more intimate nature have not quite been satisfied by robots yet (Meghdari and Alemi, 2018; Veruggio 2005).

Contemporary theories and studies of economics have turned behavioral. Behavioral Economics revolutionized mainstream neo-classical economics in the past two decades. Laboratory experiments have captured heuristics as mental short-cuts easing choices of mentally constrained human in a complex world. At the same time, heuristics were examined as a source of downfalls on rational and socially-wise choices given future uncertainty. Behavioral economists have recently started to nudge – and most recently wink – people into favorable decision outcomes, offering promising avenues to steer social responsibility in public affairs. Since then two Nobel Prizes in Economics have crowned this growing field as a wide range of psychological, economic and sociological laboratory and field experiments proved human beings deviating from rational choices and standard neo-classical profit maximization axioms often failed to explain how human behave. Human beings rather use heuristics in their day-to-day decision making. These mental short cuts enable to cope with a complex world yet also often leave individuals biased and falling astray to decision making failures. What followed was the powerful extension of behavioral insights for public policy making and international development. Behavioral economists proposed to nudge and wink citizens to make better choices for them and the community around the globe. Many different applications of rational coordination followed ranging from improved organ donations, health, wealth and time management, to name a few. Starting with the beginning of the entrance

of behavioral aspects in economic analyses and intercultural differences in behavioral understandings, these days sustainability accounting and reporting as a powerful application in a truly interdisciplinary fashion. Reporting innovatively apply behavioral economics in the professional domain. The application of behavioral economics to corporate sector economic analysis is a cutting-edge approach to capture the power of real-world relevant economics. Drawing from a line of research on bounded rationality, reporting can improve corporate success based on economic analysis tools. Delineating the potential of behavioral economics to implement market value portrays economics as a real-world relevant means to maximize value in a constantly transitioning world economy.

As one of the newest trends in Behavioral Economics, governments and institutions around the world nowadays apply behavioral economic models (Sunstein 2013) for choice architecture on regulation. In the next section it will be further analyzed how that choice architecture offers opportunities to nudge institutional and private investors into the preferred solution investments considering common sustainable criteria's and standards.

Artificial Intelligence Evolution

The human perception of and interaction with robot machines with a higher quality physical appearance differs from interaction with a computer, cell phone, or other smart devices. For robotics technology to be successful in a human-driven environment, robots do not only need to meet a level of strength, robustness, physical skills, and improved cognitive ability based on intelligence but should also fulfill a social impetus and ethical conscientiousness. The design and construction of social robots faces many challenges, one of the most important is to build robots that can comply with the needs and expectations of the human mind with cognitive capabilities coupled with social warmth. While we have Social-Cognitive Robotics (SCR) as a transdisciplinary area of research and a basis for the human-centered design of technology-oriented systems to improve human knowledge functions, judgements and decision making, collaborations, and learning; hardly any information exists on socio-evolutionary comparisons Social cognitive robotics has been evolving and verified through a series

of projects to develop advanced and modern technology-based systems to support learnings and knowledge functions, and is beginning to play an effective role in societies across the globe SCR or Socio-Cognitive Robotics is the interdisciplinary study and application of robots that are able to teach, learn and reason about how to behave in a complex world. Social robotics technology promises a many benefit but also challenges that society must be ready to confront with legal means and ethical imperatives.

Artificial Intelligence Ethics

Ethics describes moral principles that govern a person's or group's behavior. Roboethics describes the ethics and morals of robotics, the science of robots. Roboethics therefore captures the integration of ethics into AI and algorithms. So, it is not the ethics of robots or artificial ethics but the human ethics of the robot's designer, manufacturers and users. This field recently gained considerable attention among humanities and robotics engineers who draw on insights from computer science, artificial intelligence, mechanics, physics, math, electronics, cybernetics, automation and control . What specifies the emergence of socio-cognitive robotics is that humanity is at the threshold of replicating an intelligent and autonomous agent. In order to enhance the ability of social robots to successfully operate in humane ways, roles and environments, they are currently upgraded to a new level of physical skills and cognitive capabilities that embrace core social concepts (Meghdari and Alemi 2018). Robotics thereby unifies two cultures, in which complex concepts – like learning, perception, decision-making, freedom, judgement, emotions, etc. – may not have the same semantic meaning for humans and machines . In the design and construction of social robots, the consideration of ethical concerns has therefore leveraged into an imperative (Lin, Abney & Bekey 2012). Human-robot (a machine with a higher physical and social ability) interactions, are somewhat different compared to other types of human-machine interactions (i.e. with a computer, cell phone, or other smart device) , It is therefore essential for researchers, scholars, and users to clearly identify, understand, and consider these differences and ethical challenges so that they can benefit from and no one gets harmed by the assistance of social robots as a powerful tool in providing modern and quality services to society.

Robots and algorithms now taking over human decision-making tasks and entering the workforce but also encroaching our private lives, currently challenges legal systems around the globe.The attribution of human legal codes to AI is one of the most groundbreaking contemporary legal and judicial innovations. Until now legal personhood has only been attached directly or indirectly to human entities (Dowell 2018). The detachment of legal personhood from human being now remains somewhat of a paradox causing an extent of "fuzziness" of the concept of personhood (Barrat 2013; Solum 1992, 1285). As AI gets bestowed with quasi-human rights, defining factors of human personhood will need to be adjusted (Dowell 2018). Human concepts, such as morality, ownership, profitability and viability will have different meaning for AI. The need for redefining AIE has therefore reached unprecedented momentum. As predicted trend, the co-existence of AI with the human species is believed to change the fundamental concepts of social, political and legal systems. AI has already produced legal creations and will do so even more in the near future, through its developing autonomy. In addition, the technology leading to AGI and ASI is already present, posing moral and legal dilemmas about who should control it and under what terms. The emergence of AGI and ASI will necessitate the attribution of some extent and of some type of legal personhood, bearing rights and obligations. AI will not be most probably an exact replication of human intellect behavior. "[U]ultimately, robots' autonomy raises the question of their nature in the light of the existing legal categories –of whether they should be regarded as natural persons, legal persons, animals or objects– or whether a new category should be created, with its own specific features and implications as regards the attribution of rights and duties" (Committee on Legal Affairs 2016, 5). Behavioral economists add the question whether AI and robots should be created to resemble human beings' decision making with fast thinking and fallible choices or rather be targeted at perfect rationality and slow thinking (Kahneman 2011). General conscious is strived for so that AI possesses consciousness, which it can evolve and enhance on the basis of its own critical reflection and assessment of external factors . A lower level of autonomy exists if an entity can demonstrate such consciousness at a narrow field or can

self-evolve and self-adapt to external influences, thus reaching decisions "of its own," without being conscious of its intelligence as such (Tzimas 2018).

Capacities coupled with human-like emotional features, they are attributed a legal personhood in order to ensure to be comprehended correctly and to avoid unfair treatment, towards humans as well . Artificial entities are currently gaining human or quasi-human status in the Western and Arab worlds in forming an intellectual autonomy of the entity (MacDonald 2016). For instance, in Saudi Arabia the first female robot got a citizenship in 2017 and the robot appears to have more rights than a human female in Saudi Arabia.

Taxonomy development with XBRL

Behaviorally informed tools for disclosure and transparency are selected by governments (Sunstein 2013). To use a technical standard for the exchange of information, regulators or independent institutions introduce taxonomies using flexible "transparency technology XBRL (eXtensible Business Reporting Language)". It is part of the choice architecture on regulation by governments (Sunstein 2013), which applies nudging for influencing towards a preferred option. XBRL represents an open free of charge technical standard for electronic reporting and the exchange of data (Cohen, Schiavina and Servais 2005; Mirsch, Lehrer and Jung 2017; Sunstein 2013; Weinmann, Schneider and vom Brocke 2016) and should democratize the information access between institutional and private investors. XBRL inevitably requires the usage of an adequate taxonomy (Kurt and David 2003).

The taxonomy development in the context of XBRL considering the academic literature follows the following aims:

- Offer transparent corporate information to investors, which is structured so that it becomes possible to process the information by software without the requirments to manually map or human intervention and comparable information based on country-by-country or sector analysis .
- Enable the preparers to fulfill compliance requirements set by regulators, in terms of disclosing information in accordance with local and international rules .

- Improve the financial and non-financial communication by enabling adoption of specific branch requirements of industry (banks, insurance etc.) and of business variations.

However, XBRL requires a taxonomy , as the main advantage of being able to compare can only be reached by a common used taxonomy . This is also relevant for sustainability, as without a holistic standardized approach it cannot be achieved to reach sustainable goals, as institutional and private investors would follow completely different metrics. Therefore, the aim of such a sustainability taxonomy is to provide a framework for classifying all potential assets or activities against a comprehensive set of sustainability goals –from climate change to broader environmental and social goals, including the Sustainable Development Goals. The starting point for the definition of sustainability goals are the three associated risks: physical, transition and liability risk.

Different types of finance are 1) used to finance different stages of a project or asset development (e.g. acquisition/ development, operation, refinancing) and 2) used to match varying levels of inherent risks in any investment, as this can affect ability to access different types of finance.

According to , there exists no standard way to build up a taxonomy. Taxonomies can be developed for several reasons and different approaches exist from software, knowledge and ontology development for XBRL engineering. There is a best practice release by XBRL International, the "Financial Reporting Taxonomy Architecture (FRAT)", which defines modelling rules for XBRL taxonomy development (Debreceny 2009). However, this model focuses on technical aspects of how business rules are implemented in a specific XBRL taxonomy, and aspects of software engineering are integrated within this model. From a holistic point of view, the taxonomy development process encompasses reporting elements, technical XBRL specification and testing.

Existing approaches for the methodology of the development and engineering of a taxonomy in the academic literature share a focus on the technical aspects of the taxonomy development process via engineering models . The following overview follows the objective to combine business-rule development and taxonomy development.

- In the preparatory phase, reporting elements need to be defined and the associated meta-data, including specifications of the taxonomy and its intended use.
- A building phase follows, which focus on technical considerations, application rules on the base taxonomy and the management of extensions.
- Finally, there is a maintenance and evolution phase for the management and development of the taxonomy on a continued basis.

Principles-versus rule-based Taxonomy

The development of an ethical taxonomy should also consider existing best-practice taxonomies for corporate reporting. Historically, either an inductive or deductive methodology to develop a taxonomy can also be referenced to the principles-based vs. rule-based debate in the academic literature about accounting taxonomies. The principles-based vs. rule-based debate in the U.S. was rediscussed after the Enron and WorldCom accounting scandal 2002. An intense discussion whether US GAAP should become more principles-based, as rules-based standards might give rise to "cook-book accounting", without considering a substance-over-form approach. So, if there is no discretion to the chef, the taste will always be the same. US GAAP tends to be mechanical and inflexible. Clear-cut rules have some advantages, but the risk is that this approach motivates financial engineering designed specifically to circumvent these knife-edge rules, as is very often given proof in the tax literature. According to a standard should not be seen as only principles or rule-based but should rather be regarded as more or less rule-based. According to a behavioral analysis, Nelson concludes that rules can improve the accuracy of the communication of the standard setter and reduce imprecision associated with aggressive reporting due to unawareness of existing rules (Nelson 2003). Nelson does not consider that rules increase imprecision but also enable companies to structure transactions to meet the accounting rule without following the true economic substance of the transaction. This is one of the main arguments by supporter of principles or concepts-based accounting. They point to the challenge when moving from

a rule-based to a concepts-based standard setting, as informed professional judgement and expertise for the implementation is increasingly required.

In the area of ethical taxonomies, it is important to mention that ethics concerns the study and explanation of moral beliefs, so what is right or wrong. There are in general three branches , in which ethics are differentiated. Normative ethic defines how we should live in forms of principles, which we have just explained. Applied ethics are the defined rules for specific areas such as medical ethics, bioethics or business ethics. This is like the rule-based taxonomy approach. The third branch is the meta ethics, which identify what is the general nature of morality, which will not be relevant for the process of the taxonomy development.

Research methodology and introduction to Roboadvice

The concept follows the idea of the development of a uniform classification system for artificial intelligence ethics ("AIE taxonomy"). It is essential for market participants that a common understanding of ethical standards regarding the application of artificial intelligence, labels, assets and financial products exist. In a next step market, a participant will be able to build trust by providing full transparency and precise information applying these developed ethical standards. This understanding needs to be derived from legally approved, clear, consistent, comprehensible and neutral definitions that should take into consideration existing international and regional standards, which are already applied by market participants. The application of the ethical taxonomy will also enable to provide transparency on potential chances as well as risks associated with Artificial Intelligence.

What is the research method, which is applied in this paper? In the following course of this paper artificial intelligence ethics will be defined with the term used in the academic literature of "Roboethics" based on the concept of Veruggio and Operto . Veruggio and Operto provide a roadmap with the aim to monitor roboethics from a cross-cultural interdisciplinary approach. Several authors deal with roboethics with different approaches: what we intend to derive from a roboethics, is there justice, what are conditions for a robot to be moral agent , what are fundamental differences of humans and robots.

This working paper follows the approach to analyse the ethics of those designing and using robots, and the ethics of robot use, so what is built inside the robots. For this an inductive approach is applied. The use case is the market for robo advisors in Germany. In addition to that professional standards for ethics are analyzed: NSPE Code of Ethics of Engineers, IEEE Code of Ethics, ASME Code of ethics of engineers and WPI Code of Ethics for Robotics Engineers, if it is possible to incorporate those standards applying a deductive approach into the taxonomy. The deductive method consists of a methodology that changes from the general to the specific content. The associated advantage of the deductive method is that hypotheses and expected findings are developed before the data collection ("a priori"). The underlying assumptions are often based on theoretical frameworks and therefore the subsequent analysis can be assessed as logical and focused. The inductive approach derives general statements on observations and facts. An inductive researcher considers variables and considers a fully developed prior research design consisting of a literature review, models and a set of data. The usual aim is to con-struct a new framework instead of testing existing concepts. The cornerstone of the inductive method is to set up a framework based on categorization of data. One of the main advantages of the inductive method is its flexibility and openness about alternative measures and relationships. Overall a mixed-method methodology is applied in this working paper. The reason for such a design is that the same findings are generated even with different design choices, therefore diminishing the determination of the design choice and the research conclusion. Increased variation of methods to examine a topic can lead to a more robust and generalizable set of findings. Recommendations could be provided with a greater level of detail if triangulation or a mixed-method approach were applied.

Roboadvice consists of online investment guidance and portfolio management services considering algorithms and models . The overarching principle, which deviates from non-robo advice is to eliminate or reduce human intervention and to rely only on computer programmes to identify the optimal investment strategy for each individual customer. Robo-advisors are fully automated online platforms that enable customers digital financial advice and portfolio allocation. Robo advisory process can be divided into three sub-processes: 1) initial investor screening; 2) implementation of

investment strategies; and 3) monitoring and evaluation of these strategies. Implementation of investment strategies follows customer profile, which is identified following an online questionnaire. Robo-advisors select specific assets that are commensurate with investors' individual preferences. Among the spectrum of investable assets exchange-traded funds (ETFs) are very often used asset class. Automation and passive investment strategies have an important value-added function: the elimination of internal agency conflicts that can arise between financial advisors and their customers considering Principal Agent Theory. Also, the remuneration structures of financial advisory services (both commission-based and fee-based models) can also trigger conflict interest as human advisory is very often not in the best interest of the client due to moral hazard. Robo-advisors usually allocate assets using algorithms based on mean-variance optimization. Based on modern portfolio theory, higher risk returns can be achieved by maximizing returns for a given level of risk. The variance implies the risk, so the lower the variance to the mean return the more an efficient portfolio is achieved.

- Robo-advisors undergo the same requirements regarding conduct standards as human advisory services apply to and traditional financial advisors alike. Robo-advisors have the same transparency rules in terms of costs, potential risks and limitations of their services. Despite its automatic rules the duty exists to fully and fairly disclose all information so that clients can clearly understand their investment practices and potential conflicts of interest. This needs to be understandable for an independent third party, who is not an expert in robo-advice.
- Secondly, robo-advisors need to give clear evidence how they handle operational and market risk both in normal times and in distressed market conditions. Investors must be informed about operational aspects of their services, i.e. regarding the assumptions and limitations of the optimization algorithm for portfolio allocation and rebalancing.
- Thirdly, Roboadvisors should ensure that their recommendations and strategies are fit for purpose of the client's profile. Suitability should be based on the client's financial situation and investment objectives. For this, robo-advisors depend on the information provided by clients in online questionnaires. This is also circumventing ethical questions, as wrong execution or misuse of client information for not acting in the best interest would imply ethical issues.

Customer screening is one of the most crucial elements of robo-advisory. It has proven beneficial to introduce vignettes and some human touch in the form of bionic advice. Cybersecurity and the protection of sensitive customer information is a last pivotal issue when it comes to automated online advice. Thus, robo-advisors must establish controls to protect client data and to maintain the public website/the client's log-in functionality.

As Roboadvice is a fast growing business area, regulators and policymakers, as unique business models and limited or no human interaction require some clarification in certain cases . In the US, to inform robo-advisory clients, the Securities and Exchange Commission (SEC) recently published a guidance report. The SEC emphasizes that, as registered investment advisors, robo-advisors are subject to the same requirements of the Advisers Act of 1940 as non robo-advisors. In a same manner, joint committee of the three European Supervisory Authorities (ESA) launched an assessment of robo-advice, aimed at gauging whether any action was required to harness its potential benefits and mitigate its risks. End of 2016, the ESA committee decided to continue monitoring robo-advisory services, but not to apply cross-sectoral regulatory or supervisory action. Digital advice services are subject to the same regulatory requirements as traditional financial advisors and are therefore supervised by similar authorities as traditional financial advisors, i.e. the SEC and FINRA in the US, the FCA in the UK, BaFin in Germany and AMF in France.

Robo advisor market in Germany

Robo advisor market in Germany can be differentiated along three basic types considering Finanztest 2017.

Type 1: Roboadvisors solely focus on providing information how to find for customers the best product. Those type 1 act as disintermediation, as companies following this business model do not take responsibility for the investment of the clients but simply provide more transparency to the yet rather new market and new market participants. Examples of such companies are JustETF or Moneyfilter in the Germany market.

Type 2: Roboadvisors follows the business model of a passive fund management strategy. Asset management is executed based on the customer

preferences, however no active portfolio selection is performed by the robo advisor. Examples of such offerings are vaamo, easyfolio, fintegro or growney.

Type 3: Roboadvisors apply an active fund management strategy, which includes the whole asset management cycle. Examples for such product characteristics are Scalable Capital, Liquid or Quirion.

Based on a study from Oliver Wyman about 40 start-ups are in the German market, while the assets under management could increase by 2020 from currently €100 million to €30 billion by 2020, but €440 billion is expected for the global market volume of robo advisor.

In a next step the existing robo advisors are analysed with regard to their ethical considerations.

Name	Approach	Ethical considerations	Minimum investment	Costs
Vaamo	Passive	Yes		0,79%
Scalable capital	Active	Yes	t€10	0,75%
Quirion	Active	Yes	t€5	0%
Fintego	Passive	Yes	t€2.5	0,75
Whitebox	Active	Yes	t€5	0,95%

Existing Professional Standards: National Society of Professional Engineers (NSPE), Institute of Electrical and Electronic Engineers (IEEE), American Society for Mechanical Engineers (ASME), code for robotics engineers (WPI)

Professional ethics reflect standards on the interaction between professionals. As this working paper assumes that it is not the ethics of robots or artificial ethics but the human ethics of the robot's designer, manufacturers and users, the focus is on existing standards of user manufacturer of robots.

National Society of Professional Engineers (NSPE)

Based on the ethics standards of the NSPE, the following guidelines are provided.

1. To accept responsibility in making decisions consistent with the safety, health and welfare of the public, and to disclose promptly factors that might endanger the public or the environment
2. To avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist
3. To be honest and realistic in stating claims or estimates based on available data
4. To reject bribery in all its forms
5. To improve the understanding of technology, its appropriate application, and potential consequences
6. To maintain and improve our technical competence and to undertake technological tasks for other only if qualified by training or experience, or after full disclosure of pertinent limitations
7. To seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, and to credit properly the contributions of others
8. To treat fairly all persons regardless of such factors as race, religion, gender, disability, age, or national origin
9. To avoid injuring others, their property, reputation, or employment by false or malicious action
10. To assist colleagues and co-workers in their professional development and to support them in following this code of ethics.

These are very general ethical principles, which can be applied to any professionals implementing or manufacture new products applying new technologies. It provides a good foundation for the further development of the taxonomy.

Institute of Electrical and Electronic Engineers (IEEE)

Considering the IEEE, the following rather general code of conduct is formulated:

1. Using their knowledge and skill for the enhancement of human welfare
2. Being honest and impartial, and serving with fidelity their clients (including their employers) and the public; and
3. Striving to increase the competence and prestige of the engineering profession.

American Society for Mechanical Engineers (ASME)

The following code from ASME particularly focus on ethical issues arising for mechanical engineers:

1. Engineers shall hold paramount the safety, health and welfare of the public in the performance of their professional duties.
2. Engineers shall perform services only in the areas of their competence; they shall build their professional reputation on the merit of their services and shall not compete unfairly with others.
3. Engineers shall continue their professional development throughout their careers and shall provide opportunities for the professional and ethical development of those engineers under their supervision.
4. Engineers shall act in professional matters for each employer or client as faithful agents or trustees and shall avoid conflicts of interest or the appearance of conflicts of interest.
5. Engineers shall respect the proprietary information and intellectual property rights of others, including charitable organizations and professional societies in the engineering field.
6. Engineers shall associate only with reputable persons or organizations.
7. Engineers shall issue public statements only in an objective and truthful manner and shall avoid any conduct which brings discredit upon the profession.
8. Engineers shall consider environmental impact and sustainable development in the performance of their professional duties.
9. Engineers shall not seek ethical sanction against another engineer unless there is good reason to do so under relevant codes, policies and procedures governing that engineer's ethical conduct".

Code for robotics engineers (WPI)

This code is specialized to robotics engineers and can therefore adequately address roboethics issues. "As an ethical robotics engineer, I understand that I have responsibility to keep in mind at all times the wellbeing of the following communities: Global—the good of people and the environment National—the good of the people and government of my nation and its allies Local—the good of the people and environment of affected communities

Robotics Engineers—the reputation of the profession and colleagues Customers and End-Users—the expectations of the customers and end-users Employers—the financial and reputation well-being of the company To this end and to the best of my ability I will:
1. Act in such a manner that I would be willing to accept responsibility for the actions and uses of anything in which I have a part in creating.
2. Consider and respect people's physical wellbeing and rights.
3. Not knowingly misinform, and if misinformation is spread do my best to correct it.
4. Respect and follow local, national, and international laws whenever applicable.
5. Recognize and disclose any conflicts of interest.
6. Accept and offer constructive criticism.
7. Help and assist colleagues in their professional development and in following this code".

Based on the analysis of the Roboadvisors of the sample of 5 companies and the professional ethics the following ethical taxonomy is developed.

Development of AI-ethics (Roboethics) Taxonomy

Below are described the reporting elements and the required meta data to form a taxonomy complying with XBRL requirements.

The following reporting elements define the two channel on transition and physical risks and also consider as a third source AI&robotics researchers best practice:

Roboethics/AI-Ethics Taxonomy — Transition risk:
— Risk of Operational Failure

- Safety: AI-system should be safe and secure throughout the operational lifetime and verifiably so where applicable and feasible
- Failure transparency: If an AI system causes harm, it should be possible to ascertain why and provide such transparency to the client
- Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority

- Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives and to ensure that human profiles are correctly interpreted by the machines

—Risk of Value Misalignment

- Principal-agent conflict: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications
- Human Values: AI systems should be designed and operated to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity
- Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends
- Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization

—Risk of failure due to autonomous decision making

- Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation
- Human control: Human interaction is required to control internally functionality of autonomous systems
- AI-Arms Race: An arms race in lethal autonomous weapons should be avoided
- Recursive Self-improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures

—Risk of negligence

- Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
- Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources

- Shared benefit: AI technologies should benefit and empower as many people as possible

Roboethics/AI-Ethics Taxonomy — Physical Risk

The following reporting elements define the second channel on physical risk

- Physical Risk
  - —Supply Chain Risk
    - Sales impact due supply chain risk impacted by AI-failure risk leading to distribution delays, supply shortage and high price sensitivity
    - Resource demand of dependency of natural resources leading to supply shortage and high input cost
  - —Operational Risk
    - Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.
    - Socio-economic: Access to AI leading to migration and economic disruption leading to business interruptions, political instability and social license to operate
  - —Market Risk
    - Sales impacted by ethical risk leading to interruptions at point of sale, migration conflict and risk of political disruption
    - Autonomous systems might become uncontrollable and
    - Control measures might not be effective or also done by machines due to efficiency and leading to further risk of failure

Conclusions

Globalization led to an intricate set of interactive relationships between individuals, organizations and states and to an unprecedented correlation of massive global systems causing systemic risk to increase exponential. Unprecedented global interaction possibilities have made communication more complex than ever before in history as the whole has different properties than the sum of its increasing diversified parts.

This paper in the absence of a global Artificial intelligence or roboethics framework tries to put emphasis back on decision-usefulness of the investor and develops a Taxonomy considering the transparency technology Extensible Reporting Mark-up language (XBRL). The linkage to financial stability is provided by two channels of risk capturing: physical and transition risk. The study applies a mixed-method approach. Robo advice is selected as a growing market for the application of artificial intelligence in the online portfolio management without human intervention to analyze inductively existing ethical concepts and considerations. Considering professional standards on ethics for robots' manufacturer and engineers enables to derive deductively the final AI-Ethics (Roboethics) Taxonomy.

Alongside of providing an overview of behavioral sciences with an application in the corporate domain; future research should also take a critical approach to the economic analysis of the corporation. By drawing from the historical foundations of political economy, a critical stance on behavioral sciences' use for guiding on corporate concerns could also be adopted as a heterodox spin. Behavioral Economics insights should be used for improving economic analyses to improve the accuracy and efficiency of corporate sustainability reporting. The analysis could thereby also take a heterodox economics stance in order to search for interdisciplinary improvement recommendations of the use of economics for the corporate world.

Climate risk is an increasing risk to investors due to the possible value destruction of assets. High carbon emissions incur lower risks compared to physical risks like sea-level rise, extreme weather and water shortage, which we observed in the recent summer world particularly in Europe.

Investigations should feature a broad variety of research methods and tools to conduct independent projects in a truly multi-methodological approach. Overall, all these endeavors will help gain invaluable information about the interaction of economic markets with the real-world economy with direct implications for corporate decision makers.

References

Arnold, Vicky, Jean C Bedard, Jillian R Phillips, and Steve G Sutton. 2012. "The Impact of Tagging Qualitative Financial Information on Investor Decision Making: Implications for Xbrl." *International Journal of Accounting Information Systems*.

Arwas, Andrew, and Katie. 2016. "Robo-Advice 2.0: The Next Generation." *Journal of Financial Transformation Soleil* 43: 30-36.

Asaro, Peter M. 2006. "What Should We Want from a Robot Ethic." *International Review of Information Ethics 6 (12): 9-16.*

Authors, Diff. 2018. "Autonomous Weapons: An Open Letter from Ai & Robotics Researchers." 23 (2018). Online verfügbar unter, https:// futureoflife. org/open-letter-autonomous-weapons, zuletzt geprüft am.

Benston, George J, Michael Bromwich, and Alfred Wagenhofer. 2006. "Principles-Versus Rules-Based Accounting Standards: The Fasb's Standard Setting Strategy." *Abacus* 42 (2): 165-88.

Capek, Josef. 1921. "Zur Sozialen Seite Der Modernen Kunst." ln: *Das Kunstblatt* 5 (1921): 298-99.

Cellan-Jones. 2017. "The Robot Lawyers Are Here–and They're Winning." *BBC News Technology–Web*. Çevrimiçi, http://www. bbc. com/news/ technology-41829534.

Cullen, JP. " 2018. "After Hleg: The Prospects for Sustainable Finance in Europe." *The Cambridge Yearbook of European Legal Studies*.

Debreceny, Roger, Carsten Felden, Bartosz Ochocki, Maciej Piechocki, and Michal Piechocki. 2009. "Xbrl Taxonomy Engineering." In *Xbrl for Interactive Data*, 113-27: New York: Springer.

Dosi, Giovanni, Louis Galambos, and Alfonso Gambardella. 2013. *The Third Industrial Revolution in Global Business*. Cambridge University Press.

Douglas, John L, Reuben Grinberg. 2016. "Old Wine in New Bottles: Bank Investments in Fintech Companies." *Rev. Banking and Fin. L* 36:667, 84.

Duffy, Brian R. 2006. "Fundamental Issues in Social Robotics." *International Review of Information Ethics* 6 (12): 2006.

Durch, Bessere Beratung. 2018. "Digitalisierung?!"

Eric Heymann, Kevin Koerner, Marc Schattenberg. 2017. "Digital Economics: How Ai and Robotics Are Changing Our Work and Our Lives." *Deutsche Bank Research*.

Faloon, Michael, and Bernd Scherer. 2017. "Individualization of Robo-Advice." *The Journal of Wealth Management* 20(1): 30-36.

Gigerenzer, Gerd. 1999. *Simple Heuristics That Make Us Smart*. Edited by Peter M. Todd. New York: Oxford University Press. https://aalto.finna.fi/Record/alli.266138.

Financial Reporting Taxonomies Architecture 1.0. Technical report, *XBRL*, 2006, 15/5/2015, xbrl.org.

Hasperué, Waldo. 2015. "The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World." *Journal of Computer Science and Technology* 15 (02): 157-58.

Healy, Paul M, and Krishna G Palepu. 2003. "The Fall of Enron." *The Journal of Economic Perspectives* 17 (2): 3-26.

Ingram, Brandon, Daniel Jones, Andrew Lewis, Matthew Richards, Charles Rich, and Lance Schachterle. "2010. A Code of Ethics for Robotics Engineers." Paper presented at the Proceedings of the *5th ACM/IEEE International Conference on Human-robot Interaction*.

Johnson, R Burke, Anthony J Onwuegbuzie, and Lisa A Turner. 2007. "Toward a Definition of Mixed Methods Research." *Journal of mixed methods research* 1 (2): 112-33.

Kaya, Orçun, Jan Schildbach, Deutsche Bank AG, and Stefan Schneider. 2017. "Robo-Advice–a True Innovation in Asset Management." *Deutsche Bank Research*, August, available at https://www. dbresearch. com/PROD/DBR_INTERNET_EN-PROD/PROD0000000000449010/Robo-advice_-_a_true_innovation_in_asset_managemen. pdf.

Kemp, Richard. 2016. "Fourth Industrial Revolution." *The Lawyer* 31 (21): 12.

Kennedy, Gavin. 2009. "Adam Smith and the Invisible Hand: From Metaphor to Myth." *Econ Journal Watch* 6 (2): 239-263.

Kowert, Weston. 2017. "The Foreseeability of Human-Artificial Intelligence Interactions." *Texas Law Review* 96: 181.

Larson, David Allen. 2010. "Artificial Intelligence: Robots, Avatars, and the Demise of the Human Mediator." *Ohio State Journal On Dispute Resolution* 25:(1): 105.

Maines, Laureen A, Eli Bartov, Patricia Fairfield, D Eric Hirst, Teresa E Iannaconi, Russell Mallett, Catherine M Schrand, Douglas J Skinner, and Linda Vincent. 2003. "Evaluating Concepts-Based Vs. Rules-Based Approaches to Standard Setting." *Accounting Horizons* 17 (1): 73-89.

McCarthy, John. 2007. "From Here to Human-Level Ai." *Artificial Intelligence* 171 (18): 1174-82.

Meghdari, Ali, and Minoo Alemi. 2018. "Recent Advances in Social & Cognitive Robotics and Imminent Ethical Challenges." Paper in the *Proceedings of the 10th International RAIS Conference on Social Sciences and Humanities*.

Mitcham, Carl. 2005. "National Society of Professional Engineers (NSPE) Code of Ethics." *Encyclopedia of Science, Technology, and Ethics* Vol. 4. Detroit: Macmillan Reference USA.

Moore, Gordon. *1965*. "Moore's Law." *Electronics Magazine 38 (8): 114*.

Moudud-Ul-Huq, Syed. 2014. "The Role of Artificial Intelligence in the Development of Accounting Systems: A Review." *IUP Journal of Accounting Research & Audit Practices* 13, no. 2 (2014).

Nelson, Mark W. 2003. "Behavioral Evidence on the Effects of Principles-and Rules-Based Standards." *Accounting Horizons* 17 (1): 91-104.

Nobes, Christopher W. 2005. "Rules-Based Standards and the Lack of Principles in Accounting." *Accounting Horizons* 19(1): 25-34.

Parfet, William U. 2000. "Accounting Subjectivity and Earnings Management: A Preparer Perspective." *Accounting Horizons* 14 (4): 481-88.

Picard, RW. 1997. "Affective Computing Mit Press Cambridge." MA Google Scholar.

Piechocki, M., Felden, C., Gräning, A., Debreceny, R. 2009. "Design and Standardisation of Xbrl Solutions for Governance and Transparency." *International Journal of Disclosure and Governance* 6 (3): 224–40.

Piechocki, Maciej, and Carsten Felden. 2007. "Xbrl Taxonomy Engineering. Definition of Xbrl Taxonomy Development Process Model." *ECIS*.

Puaschunder, Julia M. 2018. "Artificial Intelligence Ethik (Artificial Intelligence Ethics)."

Puaschunder, Julia M. 2017. "Financing Climate Justice through Climate Change Bonds." *Oxford Journal of Finance and Risk Perspectives* 6 (3): 1-10.

Roe, S.K., and A.R. Thomas. 2013. *The Thesaurus: Review, Renaissance, and Revision*. Taylor & Francis, http://books.google.de/books?id=mO_-ePB_asQC.

Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of research and development* 3 (3): 210-29.

Shiller, Robert J. 2015. *Irrational Exuberance*: Revised and Expanded Third Edition. Princeton University Press.

Shugg, W Tillar. 1995. Handbook of Electrical and Electronic Insulating Materials. Vol. 995: *IEEE press* New York.

Sullins, John. 2011. "Introduction: Open Questions in Roboethics." *Philosophy and Technology* 24, no. 3: 233.

Sunstein, Cass R. 2013. "Nudges. Gov: Behavioral Economics and Regulation."

Swanson, Zane, George Durler, and William Remington. 20017. "How Do Firms Address Multiple Taxonomy Issues?" In *New Dimensions of Business Reporting and Xbrl*, 127-46: London: Springer.

Thietart, R.A. 2001. "Doing Management Research: A Comprehensive Guide." *SAGE Publications*, https://books.google.de/books?id=EfiB0K_oFlwC.

Todorov, Alexander. 2005. "Inferences of Competence from Faces Predict Election Outcomes." *PubMed.gov* 308 (5728): 1623. http://pubmed.gov/15947187.

Trucost. 2009. "Carbon Risks and Opportunities in the S&P 500."

Tzafestas, Spyros G. 2015. *Roboethics: A Navigating Overview*, Vol. 1046: Springer.

Tzimas, Themis. 2018. "Artificial Intelligence as Global Commons and The'international Law Supremacy'principle." Paper in the *Proceedings of the 10th International RAIS Conference on Social Sciences and Humanities*.

Veruggio, G, F Operto, PM Asaro, AS Duff, JP Sullins, BR Duffy, B Becker, D Marino, and G Tamburini. "Introductory Concepts and Outline of the Book, available at https://www.academia.edu/15317692/Introductory_Concepts_and_Outline_of_the_Book.

Veruggio, Gianmarco, Fiorella Operto, and George Bekey. 2016. "Roboethics: Social and Ethical Implications." In *Springer Handbook of Robotics*, 2135-60: Springer.

Wisskirchen, Gerlind, Blandine Thibault Biacabe, Ulrich Bormann, Annemarie Muntz, Gunda Niehaus, Guillermo Jiménez Soler, and Beatrice. 2017. "Artificial Intelligence and Robotics and Their Impact on the Workplace." *IBA Global Employment Institute von Brauchitsch*.

Yin, Robert K. 2006. "Mixed Methods Research: Are the Methods Genuinely Integrated or Merely Parallel." *Research in the Schools* 13 (1): 41-47.

Scientia Moralitas International Journal of Multidisciplinary Research
ISSN 2472-5331 (Print) | ISSN 2472-5358 (Online) | Vol. 4, No. 1, 2019

Interdependence, Morality and Human-Machine Teams: The Revenge of the Dualists

W.F. Lawless, PhD
Paine College, Augusta, GA, USA
w.lawless@icloud.com

ABSTRACT: Experience teaches that appearances can mislead, that deception frequents human affairs and that even reliable people misbehave. But for social scientists, based on their idea that the convergence of concepts derived from the intuitions of individuals (observations, self-reports, interviews) about social reality determine their primary model of the rational (social) world; i.e., what humans say they see is what exists; or, words matter; or, humans act as they cognitively think. But based on these models, the social sciences have accrued so many failures across the decades in building predictive theory that a theory of teams has until now been unimaginable, including in economics where results re-labeled as irrational have won Nobel prizes but without a foundational theory. Seemingly, concepts based on the individual promote transient norms by which to judge morality; e.g., the passing fad of self-esteem; the newest fad of implicit racism; the old fad of positive thinking. And yet, irrational and biased humans in freely organized and competitive teams manage to innovate year after year. In contrast to traditional social science, the most predictive theory in all of science is the quantum theory, each prediction confirmed by new discoveries leading to further predictions and discoveries, but the dualist nature of the quantum theory renders the meaning of physical reality meaningless despite more than a century of intense debate. By ignoring meaning, we introduce to the science of teams the quantum-like dualism of interdependence where social objects co-exist in orthogonal states. To judge the ethics of Artificial Intelligence (AI), our theory of interdependence makes successful predictions and new discoveries about

human teams that account for the poor performance of interdisciplinary science teams; explain why highly interdependent teams cannot be copied; and begin to address the newly arising problem of shared context for human-machine teams.

KEY WORDS: Interdependence, teams, subadditivity

Introduction

After Copernicus proposed his counterintuitive theory for the motions of the heavenly bodies, Kant (1755-70) reasoned that human intuitions do not conform to objects, rather,

> in a manner contradictory to the senses … [let us] seek for the observed movements not in the objects of the heaven but in their observer …

Ignoring Kant, relying on the convergence of intuitions derived from simple observations (e.g., polls), two leading decision theorists, Tetlock & Gardiner (2015), concluded that forecasting "is a skill that can be cultivated," and that prediction, politics and human affairs are not inscrutable, but rather like weather forecasting where predictions are possible, rational and accurate. To demonstrate the power with their model of superforecasting, they started a public website (Tetlock & Gardner 2015). However, their first "superforecast" that Brexit would not be supported by the British electorate failed, as did their second superforecast that Trump would not be elected the next U.S. President (Lawless 2017a,b). How does this relates to ethics?

To explain the failure of Tetlock-Gardiner and other social scientists, our research on the quantum-likeness of interdependence accounts for the dual nature of human affairs first theorized by Bohr (1955). In agreement, we have found that the more certain are social scientists about the human observations of behavior (e.g., based on converging data from interviews or self-reports on ethics), the less certain becomes the information gained about the human behavior being studied, nullifying predictability (Zell & Krizan 2014). Three examples: First, despite the strong claims over decades about the importance of self-esteem for academics and work (Diener 1984), in a 30-year meta-analysis, Baumeister and colleagues (2005) found virtually no association between self-esteem and either academics or work. Second, social

scientists extol the value of standardized tests, even though the variability in their results are unacceptable for the engineering of human-machine teams (e.g., Kuncel et al. 2007 argue that the standardized GRE test scores predict the success of graduate students, but their averaged observed correlation of less than 0.30, corrected to about 0.40, squared, means that between 80-90% in the variance of a graduate student's success is unknown; for a rehash, see Kuncel & Sackett, 2018). Third, from a news report in *Science* about an HIV prevention trial for the female mates of HIV positive males (Cohen 2013):

> The women reported using PrEP 90% of the time, and their unused returns seemed to validate that figure. But when the researchers later analyzed blood levels of drugs in the women, they found that no more than 30% had evidence of anti-HIV drugs in their body at any study visit. "There was a profound discordance between what they told us, what they brought back, and what we measured," infectious disease specialist Jeanne Marrazzo said.

These examples characterize a problem with measurements in social systems that Wendt (2015, 67) described for quantum systems as "the apparent impossibility of an objective measurement." For humans, based on the evidence, the claim can be made that the dualistic nature of interdependence creates a similar measurement problem of ethical behavior in human affairs (Lawless 2017a,b).

From Plato and Aristotle to Descartes, dualism has a rich history. An early devotee, James (1892) coined the term complementarity for different parts of consciousness sharing no knowledge with other parts (p. 206), confirmed by Gazzaniga's (2011) study of split-brain patients: "the left half did not know what the right half was processing." (p. 57) "Complementarity" is the term borrowed from James by Bohr for his theory of quantum indeterminancy (Pais, 1991, p. 424). Since Bohr, Einstein and Schrodinger, the quantum model has become the most successful predictive theory ever (Weinberg 2017). James, however, eventually rejected dualism in favor of the "practical pluralistic views" in pragmatism, a rejection transformed into today's experiential monism (Stubenberg 2017) that supports the "rational" model of making decisions. The student of James, Lovejoy (1930), remained in support of dualism:

> The revolt—within the realistic provinces of philosophical opinion—against dualism, both psychophysical and epistemological, has failed. (p. 264)

Lovejoy lost his battle. Similarly, the theory of group dynamics, introduced by Lewin (1951) has become a blind alley for his model of interdependence. Jones (1998, 6), an esteemed social psychologist, greatly admired the contributions of Kurt Lewin, the founder of group dynamics (p. 21),"Lewin argued explicitly against explanations involving individual differences … ," advice rejected by current social scientists (for a review of their focus on individual differences, see Deary, 2012). Jones (1998, p. 33) agreed that interdependence was central to social life, but he also claimed that:

> useful theory has been difficult to develop … [based on] the "*bewildering complexities*" involved in the study of interdependent relations. [emphasis added]

Not resolving these "bewildering complexities" has left researchers in the social (e.g., economic, humanistic, philosophic, networks, game theory) disciplines struggling to predict the outcomes of basic interactions, exemplified by the difficulty in replicating experiments (Nosek 2015) (The problem of replication has infected the physical sciences, notably astrophysics and other fields relying on machine learning to sift through large data bases (Wild 2018). This problem is the inability to understand how a solution was derived, to replicate results, or just to cross-examine the results (Somers 2018); left them aimless (Hofman et al. 2017); and stunned by the achievements of their colleagues in the hard sciences (e.g., physics, chemistry, biology, engineering). As well, the philosophy or history of science, engaged in endless debate, has been unable to build a foundation with which to study science (Nickles 2017).

Endless debate is a clue: Even as they preside over their exquisitely predictive discipline, quantum scientists have struggled for a century over their interpretation of the quantum (Weinberg 2017). Putting that aside momentarily, the social sciences have been built atop methodological individualism (MI), the supremeness of the individual but with no theoretical value generalizable to teams (Ahdieh 2009). Yet, at this point in human history, predictability is critical, to borrow from Kuhn (1962/1970, p. 169), to find "the solved problem" for a theory of human-machine teams, otherwise their construction and ethical use, unlike designing and perfecting bridges, will be *ad hoc*. Unlike swarms, we have learned that scientific teams,

especially the best performing ones, are highly interdependent (Cummings, 2015) (The first attack ever by a swarm of drones has already occurred: "A series of mysterious attacks against the main Russian military base in Syria, including one conducted by a swarm of armed miniature drones …" (Sly, 2018). Not having a theory of teams compounds failure; e.g., ignoring the warning by Jones about the bewildering nature of interdependence, the National Science Foundation (NSF) repeatedly and blithely calls for more interdisciplinary scientific teams in the pursuit of new research. Yet, based on the work of Cummings, the National Academy of Sciences (NAS; see Cooke & Hilton, 2015) reported that interdisciplinary scientific teams were the least productive. After a public discussion with Cummings, we concluded that the poor performance of interdisciplinary scientific teams was likely caused by redundancy.

We hypothesized that redundancy should impede the positive (ethical) effects of interdependence. However, using Shannon's information theory, Conant (1976) argued that teams and organizations should minimize interdependence (mutual information); similarly, experimental social psychologists recommend that interdependence should be statistically removed to increase the replicability of an experiment (Kenny et al. 1998, p. 235). Contradicting our hypothesis, the National Academy of Sciences had predicted that "more hands make light work" (Cooke & Hilton 2015, Ch. 1, p. 13); and Centola & Macy (2007) had predicted that social networks became more efficient as redundancy increased. But this advice from social scientists has led them along with Jones to discount the value of interdependence, analogous to believing that the study of the atom would be easier without having to deal with its "pesky" quantum effects.

In contrast, we have found that redundancy decreases interdependence (Lawless 2017a); increases the opportunity for corruption and unethical behavior (Lawless 2017b); and reduces the ability of teams to innovate (Lawless 2018). "Redundancy" is the tale of an unexpected discovery in social science based on our theory of interdependence for teams that provides mathematical metrics for human and human-machine teams. It is the first successful prediction made by our theory of interdependence (Lawless 2017a), subsequently replicated (Lawless 2017b) and leading to

new predictions and preliminary support for the work-in-progress briefly described later in this report (Lawless 2018).

In support, Cummings (2015) found that the most productive science teams maximize interdependence. From Wendt (2015), "humans live in highly interdependent societies (p. 150) … [where they form] organized, structured totalities in which parts and whole are dynamically interdependent …" (p. 134).

Interdependence transmits the dualism of constructive and destructive interference (Lawless, 2017a,b). It is the social resource available to every society to innovate and evolve ethically (Lawless,2018). Typical of societies that evolve less are autocratic countries (e.g., Cuba, Venezuela, North Korea), corrupt, unethical countries (Russia, Iran, Turkey), or both (e.g., China). Interdependence signifies a communication between two or more agents, where the interdependence inherent in public competition, such as public debate, includes the constructive or destructive signals communicated to an audience of witnesses; e.g., politics, the practice of science, juries, entertainment. The interference transmitted by interdependence derives from the competition inherent in the checks and balances that limit power or unethical behavior, demonstrated by Justice Ginsburg's (2011) unanimous ruling rejecting the Environmental Protection Agency's (EPA) rule for CO-2 until it was made ripe by the maximum available "informed assessment of competing interests."

With National Security threats arising from hypersonic missiles; modernized nuclear weapons; and the advent of human-machine teams, the motivation for faster decision-making based on a shared context is increasing (Lawless et al., 2019). How human-machine teams construct context is increasingly important, brought to the fore by the Uber self-driving car accident in Arizona that killed a pedestrian in 2018. Unlike a toy ethics problem, after reviewing this accident, the National Transportation Safety Board (NTSB 2018) reported that the car saw the pedestrian 6s early; selected its emergency brakes 1.3s early; but the brakes had been made inoperable by Uber engineers to improve the car's handling. In contrast, the human operator saw the pedestrian 1s early and hit the brakes 1s after impact. The car performed as designed and faster than the human who performed

much slower. However, the car did not alert the driver of the change in context although it could have done so earlier, maybe in time to save the pedestrian; by not contributing to the context shared by the human-machine team, the car was a poor team player, a problem that we AI scientists must argue that can and should be fixed to improve social welfare.

Needed is a theory of teams like ours modeled by the mathematics of interdependence to build a shared context that we continue to develop and briefly review herein. Without a mathematics of interdependence, human-machine teams will remain ad hoc, inefficient or not effective; ethically, their contribution to social welfare may be poor. For a mathematical grasp of interdependence, which works like quantum entanglement, we have divided its effects into bistable views (e.g., action-observation; Tribe-1 versus Tribe-2; prosecutors versus defense attorneys; Einstein's interpretation of reality versus Bohr's); a measurement problem where the convergence of interpretations into the supposedly "ethical" one produces incompleteness and uncertainty by dismissing the rejected alternative interpretation (e.g., despite their lack of validity, thereby increasing the value of static questionnaires that falsely associated an individual's performance with "self-esteem," in Baumeister et al. 2005; implicit racism, in Blanton et al. 2009; or positive thoughts; in Diener 1984); and the inability to factor social states (e.g., the measurement of an interdependent social object affects the behavior and cognitions of the objects measured).

As a simple example of the effects of interdependence. Elk overgraze in forests without the presence of coyotes, making their forests unhealthy; in contrast, elk in forests where wolves have been introduced take a bite of grass and scan about in their vigilance for wolves, take another bite of grass and scan about again, continuously eating and surveilling, the intermittent eating producing a greener and healthier forest (Carroll 2016).

The bistable views of different tribes

In Kuhn's (1977) view, a set of ideas developed within a paradigm impede alternative views of reality arising between different cultures or groups (Tajfeld 1970), like liberalism versus conservatism, prosecutors versus defense

attorneys, or pre-Planckian physicists versus quantum physicists, generating tension whenever a questionable ethical event cannot be explained by groups holding different views or by the beliefs prevailing in a single group, easily dismissed when there are no means to test ideas about an ethics anomaly, but when there is, creating the tension essential to change (e.g., Martin Luther King's activism against Jim Crow laws; in Layne 2015). Unlike philosophy which is debatable but untestable, physical theories, made testable by their predictions with mathematics, create tension naturally when users consider an equation's (ethical) implications or its generalizations to establish new physical theory; still, without question, an equation's interpretations or (ethical) paradoxes derived from the predictions established by an equation can create unending conflict like with the endless quantum debates.

We have found that the value of the constructive-destructive interference transmitted by interdependence depends on the free movement of ideas, people and capital attracted by the different interpretations transmitted, why the first target of an autocratic government is to censor the interpretations it rejects (e.g., see the *New York Times* article about censorship in Turkey; in Gall, 2018; or the destruction of the entire Uighur culture by China to gain the fullest compliance of Uighurs to China's leadership; in Chin & Bürge, 2019); why a business might try to silence its opposition (NYT, 2018); or why a majority religion might choose to persecute a religious minority (Kishi, 2017). By forcibly preventing the power of teams, tribes or cultures to flourish in favor of individuals who can be more easily controlled, however, the destructive interference of censorship kills the civic, ethical, social and intellectual productivity needed for innovation (for the effects of censorship in Russia, see Varadarajan, 2018).

Innovation. One of the largest producers of patents, China, has neutralized its advantage with wide-spread censorship. Consider that the R&D expenditures by China are second in the world to the U.S. (Zumbrun, 2018). But China's state directed finance, its weak intellectual property protections and its rampant corruption impede innovation and social welfare. In China (Tamplin, 2018),

Small private-sector firms often only have access to capital through expensive shadow banking channels, and risk that some better connected, state backed firm will make off with their designs--with little recourse.

Convergence: The statistical convergence of evidence away from bistability in support of one or another concept assumes that individuals are independent; that the data generated by individuals is also independent (e.g., "detrended"); and that whatever is rejected by the evidence is noise and not a valid alternative. Social scientists look for the associations identified by correlations, accepting that while correlations do not establish causality, they believe that the lack of a correlation indicates no causal relationship. But "convergence" may not a satisfactory concept on which to base ethical judgments.

Orthogonality: If the lack of a correlation indicates no causal relationship, and the data produced is independent, a problem occurs when data is derived from orthogonal, bistable sources identified as independent by definition; e.g., husband-wife teams; CBS-Viacom businesses; pitcher-catcher role players. But after Gazzaniga's (2011) extraordinary discovery that the brains of split-brain patients produce interpretations of reality completely different for the two halves (p. 57), unexpected because no evidence of bistability existed beforehand, we assert that the lack of a correlation does not preclude causality. If members of a team are playing orthogonal roles (e.g., the bistable information arising from the different members of society's "team" in its ethical search for justice, composed of a judge, prosecutor and defense attorney), that incommensurability accounts for the negligible correlations from studies with concepts of behaviors versus actual behaviors (e.g., an alcoholic's "denial" of being an alcoholic; the different interpretations of the cause of conflict arising during a divorce; the negligible correlations between self-esteem and academic or work performance). Orthogonality accounts for the here-to-fore hidden value of teamwork; namely, humans are poor at multitasking (Wickens, 1992), but multitasking by teammates playing orthogonal roles while working toward a common (ethical) goal is the function of teams (Lawless, 2017a,b).

Measurement problem

When bistable views are censored for whatever reason, a measurement problem occurs. We have argued that censorship converts teams or tribes into a collection of individuals, allowing us to apply Shannon information theory to the result. From Shannon, in words, joint information is greater than (as the dependence between agents increases) or equal (as the independence between agents increases) to the information from its contributors (Where is the joint entropy of two sources or two agents and or is the entropy of one agent, giving):

$$information_{joint} \geq information_{agent1}, information_{agent2} \quad (1)$$

(mathematically, $H_{A,B}$ is the joint entropy by two independent sources or agents and H_A or H_B is the entropy of a single agent, giving: $H_{A,B} \geq H_{A,}H_B$). By applying Conant (1976) to Shannon information, censorship reduces the value of interdependence as a resource. In Shannon's model, deception has little to no biological value; but in biology and with humans, deceptive (and unethical) behavior serves a critical function (Chagnon, 1988), especially under authoritarian governments.

In contrast to Ginsburg's (2011) "informed assessment of competing interests," few appreciate that the value of bistability transmitted by interdependence improves social welfare by solving (ethical) problems (Kuhn, 1962/1970). Instead, recently, there has been a turn away from the bistability inherent in "checks and balances" as a means to improve social welfare by replacing it with (Vermeule, 2018):

> the administrative state ... [where its] agents may have a great deal of discretion to further human dignity and the common good, defined entirely in substantive rather than procedural-technical terms. ... agents with administrative control over default rules may nudge whole populations in desirable directions, in an exercise of "soft paternalism" ...

Vermeule's hopes are wistful. In addition to the "soft paternalism" exhibited above by EPA rejected by Justice Ginsburg (2011), or the censorship promoted inside of Turkey or Russia, the turn away from checks and balances

threatens social welfare, illustrated by the time when the U.S. Department of Energy (DOE) operated almost unimpeded by public oversight, a time when DOE alone had the authority implied by the "soft paternalism" in its management of military nuclear wastes for the "common good" when, instead, DOE's single-mindedness produced extraordinary contamination of the environment across the U.S. (Lawless et al., 2014). Further, in the cleanup since, motivated by DOE's guidance to use the cooperation inherent in consensus-seeking for decision-making by DOE's Citizens Advisory Boards (CAB), its CAB at Hanford provides a comparison versus the bistability inherent in the majority-ruled CAB at DOE's Savannah River Site in SC, one of the sites which rejected consensus-seeking in favor of majority rules to make its decisions. The result: SRS has had a significantly better, faster and safer cleanup than the Hanford site, the latter mired in endless debate and legal strife; e.g., even though the process for the vitrification of high-level radioactive wastes was innovated at Hanford, vitrification began at SRS in 1996 but has not yet begun at Hanford, and may not start there for another decade if ever. As we had predicted, and as supported by the European Union (WP, 2001), consensus-seeking is how a minority censors or controls a majority by blocking its ability to make a decision:

> The requirement for consensus in the European Council often holds policy-making hostage to national interests in areas which Council could and should decide by a qualified majority. (p. 29)

Non-factorability

Applying Von Neumann's model of constructive and destructive interference to a state of interdependence, the joint information becomes less than (as the teamwork increases between agents) or equal (as the teamwork between agents ceases, becoming equal to Shannon information) to its contributors:

$$information_{joint} \leq information_{agent1} + information_{agent2} \quad (2)$$

(mathematically, $S_{A,B}$ is the joint entropy of two interdependent sources or agents and S_A or S_B is the entropy of one agent, giving: $S_{A,B} \leq S_{A,} + S_B$). Equation (2) accounts for non-factorability. Mindful of Kant, it confirms

the biological value of how deception is applied by "fitting in," including for humans (a con artist; a military feint; a private affair). But, more importantly, Equation (2) predicts that when a team is working to perfection, the information it generates disappears as the information from its interactions go dark, meaning that the effect of counting the contributions from a team's members by an outside observer is no longer trustworthy (viz., by reducing the degrees of freedom in a team as a team begins to operate as a "unit"; in Lawless 2017b). This result explains why the performance of a perfect team is difficult or impossible to copy, even by the perfect team itself (The inability to copy interdependence is similar to the "no cloning" rule in quantum information theory (Wooters & Zurek 2009, 77). It also explains why a coach or a leader for the best teams is often necessary, inadvertently making the "best" coaches invaluable.

Future research

To advance previous research (Lawless 2017a,b), the plan is to introduce the value of intelligence as a tool used by teams to manage interdependence. Here's the problem: Although game theory was introduced in social psychology by Thibaut & Kelley (1959), Kelley (1979) abandoned it after realizing that no matter the strength of preferences chosen on paper by subjects before playing a game, subjects were too responsive to the interdependent feedback from the choices made by their opponents during actual games (Lawless, 2017a). Kelly abandoned game theory for close relationship theory, but correlations for that theory also failed to establish the value of interdependence for two similar reasons: First, matching two people for a relationship based on choices selected with a piece of paper is again overwhelmed by the (constructive or destructive) interference from interdependence inherent in a partnership. Second, interdependence theory indicates that the best relationships are those built around partners in orthogonal roles (e.g., all else equal, instead of the inferior performance derived from the destructive interference caused by two catchers and one pitcher playing in a baseball game simultaneously, a better arrangement is the constructive interference from only one catcher, one pitcher and one first baseman playing together at a time). Orthogonal

information, however, produces zero correlations. Intelligence enters when making the partnership choices that minimize destructive and maximize constructive interference among partners and only when the partners agree to a superordinate (ethical) goal for their team.

The prevalence in the social interaction of interdependence forces social navigators to rely on intelligence during a competition to craft a social path that achieves a team's superordinate goal (mission, ethical behavior) by amplifying its skills with constructive interference, mindfully using destructive interference to sharpen its focus, by deploying team boundaries to block outside interference, but thereby making its decision process opaque (zero correlations). Intelligence determines the members selected for a team (constructive); the shape of a team's structure that produces maximum entropy (MEP; see Wissner-Gross & Freer 2013; i.e., maximum work output for a team of workers; or maximum exploration of a solution space for a team of scientists); and the shortest social path with MEP to overcome obstacles (Martyushev 2013) to achieve a team's superordinate goal to guide and measure its progress (Lawless 2018). The quantum-like nature of interdependence causes tension between the intuitions leaders use in tradeoffs under uncertainty that shape a team and its structure to achieve MEP (e.g., to maximize performance, leaders choose the skills a team needs in its competitions, their internal communications, and the configuration of the structure that shapes the configuration of its members; in England, 2013).

Conclusions

In a free society, because of the costs of extras, interdependence automatically reduces redundancy. While the meaning of interdependence is meaningless (Jones' *bewilderment*), we conclude that interdependence is the primary resource free societies harness to shape their teams and structures to improve (ethical) social welfare. That is the reason authoritarians attempt to quash interdependence as their first order of business (by censoring free speech; by ending the freedom to assemble; by preventing the free exercise of religion; by destroying competing cultures; etc.).

Interdependence is the science of human and human-machine teams, organizations and societies; it lends itself to mathematical models, to trial and error tradeoffs, but not to a single interpretation; it could rehabilitate the social sciences and, with Kant, the integration of the philosophy of science with the history of science (e.g., Nickles 2017). It is a social science that offers interdisciplinary teams the opportunity to contribute when their skills are demanded to complete a team, but not for the specious purpose of satisfying the bureaucratic whims of an agency like NSF or DOE. It is the science of dualism with social people, organisms and future robots working with humans to build human-machine teams. Finally, to end the interminable quantum debates, Weinberg (2017) wants quantum theory to be revised so it does not give a status to human observers; good luck with that!

Interdependence is not a silver bullet. It is a trial and error approach to selecting the best members of a team with the least redundancy possible. Choosing the best teammates possible is critical. Training is essential. Supporting players by offering rest or providing relief or substitution is necessary (e.g., even the best teams aboard Navy ships need relief after 8-hour or longer shifts; in Holmes, 2018). But the extraordinary value of interdependence to societies also helps to reduce the alarm from armies of robot slaves: armies of slaves will be no match for intelligent teams operating at maximum performance.

MI theories are neither foundational nor do they afford additive building blocks; unlike MI, interdependence advances social theory. It will lead to more ethical behavior and better judgments of what is, or is not, ethical. At a minimum, AI must provide a context that both humans and machines can share and come to trust, unlike the Uber car that did not share its context with its human operator. That way, when there are rules to follow that society has helped to establish (e.g., guided by Justice Ginsburg, 2011), human-machine teams will be able to use their intelligence to abide by the (ethical) rules and complete their mission.

In closing, Wendt (2015, p. 34) adds that a quantum-like model "offers the potential for revealing new social phenomena", which we have demonstrated by establishing the value of team boundaries, the multitasking nature of teams, and the size of teams, heretofore an open problem (Cooke &

Hilton 2015, 33); e.g., for the latter, to wit, in agreement with the second law of thermodynamics, the smallest size of a perfect team is one that minimizes its redundancy, maximizes its interdependence and yet still manages to complete its mission (Lawless 2017a,b).

Acknowledgements: An earlier version of this paper was presented at *the 9th International RAIS Conference on Social Sciences and Humanities*, at Princeton, The Erdman Center, USA,on April 4-5, 2018.

References

Ahdieh, R.G. 2009. *Beyond individualism and economics*, retrieved 12/5/09 from ssrn.com/abstract=1518836.

Baumeister, R. F., Campbell, J.D., Krueger, J.I., & Vohs, K.D. 2005, January. "Exploding the self-esteem myth." *Scientific American, 292(1)*: 84-91; from https://www.uvm.edu/~wgibson/PDF/Self-Esteem%20Myth.pdf.

Blanton, Hart, Klick, J., Mitchell, G., Jaccard, J., Mellers, B. & Tetlock, P.E. 2009. "Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT." *Journal of Applied Psychology 94*(3): 567–582.

Bohr, N. 1955. "Science and the unity of knowledge." In L. Leary (ed.), *The unity of knowledge*, pp. 44-62, New York: Doubleday.

Carroll, S. 2016. *The big picture. On the Origins of Life, Meaning, and the Universe Itself.* New York, NY: Dutton (Penguin Random House).

Centola, D. & Macy, M. 2007. "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology*, 113(3): 702–34.

Chagnon, N. 1988. "Life Histories, Blood Revenge, and Warfare in a Tribal Population." *Science* 239: 985–92.

Chin, J. & Bürge, C. 2019, 3/20."After Mass Detentions, China Razes Muslim Communities to Build a Loyal City. Authorities take down once-bustling Uighur neighborhoods to create a compliant economic hub," *Wall Street a Journal*, from https://www.wsj.com/articles/after-mass-detentions-china-razes-muslim-communities-to-build-a-loyal-city-11553133870.

Cohen, J. 2013. Human Nature Sinks HIV Prevention Trial, *Science*, 351: 1160, from http://www.sciencemag.org/news/2013/03/human-nature-sinks-hiv-prevention-trial

Conant, R. C. 1976. "Laws of information which govern systems." I*EEE Transaction on Systems, Man, and Cybernetics* **6**: 240-255.

Cooke, N.J. & Hilton, M.L. (Eds.) 2015. Enhancing the Effectiveness of Team Science. Authors: Committee on the Science of Team Science; Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education; National Research Council. Washington (DC): National Academies Press.

Cummings, J. 2015. Team Science Successes and Challenges. National Science Foundation Sponsored Workshop on Fundamentals of Team Science and the Science of Team Science (June 2), Bethesda MD (https://www.ohsu.edu/xd/education/schools/school-of-medicine/departments/clinical-departments/radiation-medicine/upload/12-_cummings_talk.pdf).

Deary, I.J. 2012. "Individual differences in human intelligence are of interest to a wide range of psychologists and to many people outside the discipline." *Annual Review of Psychology* 63:453-482, from https://doi.org/10.1146/annurev-psych-120710-100353.

Diener, E. 1984. "Subjective well-being." *Psychological Bulletin* 95(3): 542-575.

England, J.L. 2013. "Statistical physics of self-replication." *J. Chem. Phys.* 139, 121923 (2013), doi: 10.1063/1.4818538.

Gall, C. 2018, 3/4. "Erdogan's Next Target as He Restricts Turkey's Democracy: The Internet." *New York Times*, from https://www.nytimes.com/2018/03/04/world/europe/turkey-erdogan-internet-law-restrictions.html.

Gazzaniga, M.S. 2011. *Who's in charge? Free will and the science of the brain*. New York: Ecco.

Ginsburg, R.B. 2011. American Electric Power Co., INC., Et Al. v. Connecticut Et Al., 10-174, http://www.supremecourt.gov/opinions/10pdf/10-174.pdf.

Hofman, J.M. and Sharma, A. and Watts, D.J. 2017. "Prediction and explanation in social systems." *Science* 355: 486–488.

Holmes, J. 2018, 4/2. "Break and Remake the U.S. Navy Surface Fleet. Why did it take a series of shipboard disasters to jolt the U.S. Navy into reforming the training regimen?" *RCD*, from https://www.realcleardefense.com/articles/2018/04/02/break_and_remake_the_us_navy_surface_fleet_113271.html.

James, W. 1892/1950. The Principles of Psychology, 2 vols. (1890). Dover Publications. (p. 206).

Jones, E.E. 1998. "Major developments in five decades of social psychology." In Gilbert, D.T., Fiske, S.T., & Lindzey, G., *The Handbook of Social Psychology*, Vol. I, pp. 3-57. Boston: McGraw-Hill.

Kant, I. 1755-1770. *Critique of Pure Reason* (The Cambridge Edition of the Works of Immanuel Kant, 1998), Paul Guyer & Allen W. Wood (Ed., Tr.). Cambridge, UK: Cambridge University Press.

Kelley, H.H. 1979. *Personal relationships: Their structure and processes*. Hillsdale, NJ: Lawrence Earlbaum.

Kenny, D. A., Kashy, D.A., & Bolger, N. 1998. Data analyses in social psychology. Handbook of Social Psychology. D. T. Gilbert, Fiske, S.T. & Lindzey, G. . Boston, MA, McGraw-Hill. 4th Ed., Vol. 1: pp. 233-65.

Kishi, K. 2017, 6/9. "Christians faced widespread harassment in 2015, but mostly in Christian-majority countries", *Pew Research Center,* from http://www.pewresearch.org/fact-tank/2017/06/09/christians-faced-widespread-harassment-in-2015-but-mostly-in-christian-majority-countries/

Kuhn, Thomas S., [1962] 1970a. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press. Second edition with added Postscript—1969) published in 1970.

Kuhn, T. 1977. "The essential tension." University of Chicago Press.

Kuncel, N. R. & Hezlett, S. A. 2007. "Standardized tests predict graduate students' success." *Science 315*(5815), 1080-1081, doi: http://dx.doi.org/10.1126/science.1136618.

Kuncel, N. & Sackett, P. 2018, 3/8. "The Truth About the SAT and ACT. Myths abound about standardized tests, but the research is clear: They provide an invaluable measure of how students are likely to perform in college and beyond." *Wall Street Journal,* from https://www.wsj.com/articles/the-truth-about-the-sat-and-act-1520521861

Lawless, W.F., Akiyoshi, Mito, Angjellari-Dajcic, Fiorentina & Whitton, John. 2014. "Public consent for the geologic disposal of highly radioactive wastes and spent nuclear fuel." *International Journal of Environmental Studies* 71(1): 41-62.

Lawless, W.F. 2017a. "The entangled nature of interdependence. Bistability, irreproducibility and uncertainty." *Journal of Mathematical Psychology*, 78: 51-64.

Lawless, W.F. 2017b. "The physics of teams: Interdependence, measurable entropy and computational emotion." *Frontiers physics* 5:30. doi: 10.3389/fphy.2017.00030.

Lawless, W.F. 2018. "The mathematics of interdependence for superordinate decision-making with teams." *15th International Conference on Distributed Computing and Artificial Intelligence* (DECON'2018), Toledo, Spain, 20-22 June.

Lawless, W.F., Mittu, R., Sofge, D.A. & Hiatt, L. 2019; in press. Introduction to the Special Issue, "Artificial intelligence (AI), autonomy and human-machine teams: Interdependence, context and explainable AI," *AI Magazine*.

Layne, R. 2015, Fall. Fighting for Freedom with Martin Luther King Jr., The Stanford Freedom Project ~ Informed opinions through history, literature, philosophy, and contemporary experience, from https://stanfordfreedomproject.com/what-is-freedom-new-essays-fall-2014/fighting-for-freedom-with-martin-luther-king-jr/.

Lewin, K. 1951. *Field theory of social science. Selected theoretical papers*. Darwin Cartwright (Ed.). New York: Harper & Brothers.

Lovejoy, A.O. 1930. *The Revolt against Dualism: An Inquiry Concerning The Existence Of Ideas*. London: George Allen & Unwin Ltd.

Martyushev, L.M. 2013. "Entropy and entropy production: Old misconceptions and new breakthroughs." *Entropy* 15: 1152-70.

Nickles, Thomas. 2017, Summer. "Historicist Theories of Scientific Rationality." *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), from https://plato.stanford.edu/archives/sum2017/entries/rationality-historicist/.

Nosek, B., corresponding author from OCS. 2015. "Open Collaboration of Science: Estimating the reproducibility of psychological science." *Science* 349 (6251): 943; supplementary: 4716-1 to 4716-9.

NTSB. 2018, 5/24. "Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle." *National Transportation Safety Board*, from https://www.ntsb.gov/news/press-releases/Pages/NR20180524.aspx.

NYT. 2018, 3/3. "Honduras Police Arrest Executive in Killing of Berta Cáceres, Indigenous Activist." *New York Times*, from https://www.nytimes.com/2018/03/03/world/americas/honduras-berta-caceres.html.

Pais, A. 1991. *Niels Bohr's Times: In Physics, Philosophy, and Polity*. Oxford, UK: Clarendon Press.

Sly, L. 2018, 1/10. "Who is attacking Russia's bases in Syria? A new mystery emerges in the war." *Washington Post*, from https://www.washingtonpost.com/world/who-is-attacking-russias-main-base-in-syria-a-new-mystery-emerges-in-the-war/2018/01/09/4fdaea70-f48d-11e7-9af7-a50bc3300042_story.html?utm_term=.cb338c653ed8.

Somers, J. 2018, 4/5. "The Scientific Paper Is Obsolete." *The Atlantic*, from https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/.

Stubenberg, Leopold. 2017. "Neutral Monism." *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), from https://plato.stanford.edu/archives/win2017/entries/neutral-monism/.

Tajfel, H. 1970. "Experiments in intergroup discrimination." *Scientific American* 223(2): 96-102.

Taplin, N. 2018, 5/14. "Can China's red capital really innovate? U.S. technology theft from Britain helped k2018, ick-start the industrial revolution on American shores. Will China be able to replicate that success?" *Wall Street Journal*, 2018, From https://www.wsj.com/articles/can-chinas-red-capital-really-innovate-1526299173.

Tetlock, P.E. & Gardner, D. 2015. Superforecasting: The Art and Science of Prediction, Crown.

Tetlock's website. Available at http://goodjudgment.com/superforecasting/index.php/2016/11/03/is-donald-trump-mr-brexit/.

Thibaut, J. W.; Kelley, H. H. 1959. *The social psychology of groups*. New York: Wiley. ISBN 978-0-88738-633-6.

Varadarajan, T. 2018, 3/9. "Will Putin Ever Leave? Could He if He Wanted? A Stalin biographer contemplates Russia's weakness today, which makes its current ruler such a threat to the West" *Wall Street Journal*, from https://www.wsj.com/articles/will-putin-ever-leave-could-he-if-he-wanted-1520635050.

Vermeule, A. 2018, Spring. "Integration from Within. Review Essay: *Why Liberalism Failed*, by Patrick J. Deneen, Yale University Press, 2018, *American Affairs* 2(1), from https://americanaffairsjournal.org/2018/02/integration-from-within/.

Weinberg, S. 2017, 1/19. "The Trouble with Quantum Mechanics." *The New York Review of Books*, from http://www.nybooks.com/articles/2017/01/19/trouble-with-quantum-mechanics/

Weinberg, S. 2017b, 4/6. Steven Weinberg and the Puzzle of Quantum Mechanics, replies by N. David Mermin, Jeremy Bernstein, Michael Nauenberg, Jean Bricmont, and Sheldon Goldstein, et al. In response to: The Trouble with Quantum Mechanics from the January 19, 2017 issue; *The New York Review of Books*, from http://www.nybooks.com/articles/2017/04/06/steven-weinberg-puzzle-quantum-mechanics/

Wendt, A. 2015. *Quantum Mind and Social Science. Unifying Physical and Social Ontology*. Cambridge, UK: Cambridge University Press.

Wickens, C. D. 1992. *Engineering psychology and human performance* (second edition). Columbus, OH, Merrill.

Wild, S. 2018, 4/4. "Irreproducible astronomy. A combination of data-churning telescopes and proprietary algorithms has led to a slate of opaque research that has some astronomers concerned." *Physics Today*, Research & Technology, DOI:10.1063/PT.6.1.20180404a from https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180404a/full/.

Wissner-Gross, A. D., and C. E. Freer. 2013. "Causal Entropic Forces." *Physical Review Letters* 110(168702): 1-5.

Wooters, W.K. & Zurek, W.H. 2009, Feb. The no-cloning theorem, *Physics Today*, pp. 76-77; http://www.physics.umd.edu/studinfo/courses/Phys402/AnlageSpring09/TheNoCloningTheoremWoottersPhysicsTodayFeb2009p76.pdf.

WP 2001. White Paper. European governance (COM (2001) 428 final; Brussels, 25.7.2001). Brussels, Commission of the European Community.

Zell, E. & Krizan, Z. 2014. "Do People Have Insight Into Their Abilities? A Metasynthesis?" *Perspectives on Psychological Science* 9(2): 111-125.

Zumbrun, J. 2018, 4/10. "Should the U.S. worry that China is closing in on its lead in research and development? Amid a productivity slump, the IMF sees benefits from Chinese and South Korean innovation." *Wall Street Journal*, from https://blogs.wsj.com/economics/2018/04/10/should-the-us-worry-about-china-rd/.

Scientia Moralitas International Journal of Multidisciplinary Research
ISSN 2472-5331 (Print) | ISSN 2472-5358 (Online) | Vol. 4, No. 1, 2019

Artificial Intelligence Evolution: On the virtue of killing in the artificial age

Julia M. Puaschunder*
The New School, Department of Economics,
Schwartz Center for Economic Policy Analysis,
Columbia University, Graduate School of Arts, New York,
Julia.Puaschunder@columbia.edu,
Princeton University, Julia.Puaschunder@princeton.edu,
George Washington University, jpuaschunder@gwu.edu

ABSTRACT: Artificial Intelligence (AI) poses historically unique challenges for humankind. In a world, where there is a currently ongoing blend between human beings and artificial intelligence, the emerging autonomy of AI holds unique potentials of eternal life. With AI being endowed with quasi-human rights and citizenship in the Western and Arabic worlds, the question arises how to handle overpopulation but also misbehavior of AI? Should AI become eternal or is there a virtue in switching off AI at a certain point? If so, we may have to redefine laws around killing, define a virtue of killing and draw on philosophy to answer the question how to handle the abyss of killing AI with ethical grace, rational efficiency and fair style. The presented theoretical results will set the ground for a controlled AI-evolution in the 21st century, in which humankind determines which traits should remain dominant and which are meant to be killed.
KEY WORDS: Artificial Intelligence, AI, algorithms, cognitive robotics, AI-evolution, emerging technologies, ethical issues, ethics, human robot interaction, international law, killing, legal personhood, roboethics, robot-rights, social robots, virtue of killing

* Financial support of the Research Association for Interdisciplinary Studies, The New School Dean's Office, The New School Department of Economics, The New School Eugene Lang College, The New School Fee Board, The New School for Social Research, The New School for Public Affairs, University of Vienna and Vernon Arts and Science is gratefully acknowledged. The author declares no conflict of interest. All omissions, errors and misunderstandings in this piece are solely the author's.

1. Introduction

Artificial Intelligence (AI) poses historically unique challenges for humankind. As emerging globally trend, AI is extending its presence at almost all levels of social conduct and thereby raised both – high expectations but also grave concerns (Cellan-Jones 2018; Sofge 2015; United Nations 2017). With the dramatic growth in diversity and entrance of emerging technologies in today's societies, such as social robots, lifelike computer graphics (avatars), and virtual reality tools and haptic systems, the social complexity of these challenges are on the rise (Meghdari & Alemi 2018). One of the main challenges in developing and applying modern technologies in our societies is the identification and consideration of ethical issues surrounding AI (Meghdari & Alemi, 2018). The call for AI Ethics (AIE) has emerged. A growing number of AI and robotics researchers have demanded to create a framework on AI ethics building on the benefits of humanities, philosophy, natural sciencies, sociology, and social neuroscience.

AI will hold the potential to replicate human existence but also grant eternal being opportunities. In the eye of overpopulation concerns, finding mechanisms to switch off AI would be a solution to avoid a crowding of the planet. But AI currently also reaches quasi-human status through actual personhood – e.g., via citizenship and quasi-human rights applied in the Common Law but also Roman Law territories of the US and the EU. Leveraging AI entities to the status of being through the attribution of legal personhood raises challenging legal and ethical questions. Programming AI to switch itself off or switch off AI at a certain point to curb overpopulation but also as quality control against harmful behavior arising out of AI, thereby appears critical as it would come close to suicide or killing. A novel predicament between eternity and overpopulation hence calls for revising legal codes for killing and ethical imperatives and religious concerns over suicide.

But how to argue the right to terminate AI legally? And when to pull the plug? We may want to draw on the ethics of dying and virtues of killing as well as suicide literature to answer these novel questions arising out of AI. When considering the opportunity to determine life and death

of AI, humankind will see the opportunity of AI-evolution understood as a human-made evolution determining what contents survive and what to die following the goal to improve the overall offspring and general well-being of humankind. The proposed frame will offer innovative insights for legal conduct but also overlapping generations relationships. The nature of algorithms and digital technology being global demands for an international response, potentially via international law supremacy principle (Themistoklis 2018). In this paper, the novel and multidisciplinary area of socio-cognitive robotics, and the ethical challenges of emerging technologies are explored. Key ethical features based on past and present research in a variety of AI areas will be presented.

The paper is structured as follows: First, the ontology of AI is presented as well as an analysis of legal personhood. Then, the predicament between eternal life and overpopulation is addressed. The virtues of dying and killing but also philosophical arguments for the right to live or choose suicide are discussed. The paper closes with an international law and future research prospects on regulating AI and overall future outlook.

2. Theory

2.1 Artificial Intelligence

Artificial Intelligence (AI) is "a broad set of methods, algorithms, and technologies that make software 'smart' in a way that may seem human-like to an outside observer" (Noyes 2016). The "human-like" intelligence of machines derives from machines being created to think like humans but at the same time to also act rationally (Laton, 2016; Russell & Norvig 1995; Themistoklis 2018). AI is perceived as innovative technology or as the sum of different technological advances as the privilege of the private, technological sector with little — if any — public regulation (Dowell, 2018).

As the most novel trend, AI, robots and algorithms are believed to soon disrupt the economy and employment patterns. With the advancement of technologies, employment patterns will shift to a polarization between AI's rationality and humaness. Robots and social machines have already replaced people in a variety of jobs – e.g. airports smart flight check-in kiosks or self-

check-outs instead of traditional cashiers. Almost all traditional professional are prospected to be influsedwith or influenced by AI, algorithms and robotics. For instance, robots have already begun to serve in the medical and health care profession, law and – of course – IT, transportation, retail, logistics and finance, to name a few. Social robotics may also serve as quasi-servants that overwhelmingly impact our relationships. Already, social robots are beginning to take care of our elderly and children, and some studies are currently underway on the effects of such care (Alemi, Meghdari & Saffari 2017). Not only will AI and robots offer luxuries of affordability and democratization of access to services as they will be – on the long run – commercially more affordable and readily available to serve all humanity; but also does the longeavity potential of machines outperform any human ever having lived (Hayes, 2018). However, the new technology also comes with the price of overpopulation problems and the potential for misuse and violent action. Just like many other technologies, robots could be misused for wars, terrorism, violence and opression (Alemi et al. 2017).

AI's entrance in society will revolutionize the interaction between humans and AI with amply legal, moral and social implications (Kowert, 2017; Larson 2010). Autonomous AI entities are currently on the way to become as legal quasi-human beings, hence self-rule autonomous entities (Themistoklis 2018). AI is in principle distinguished between weak AI, where "the computer is merely an instrument for investigating cognitive processes" and strong AI, where "[t]he processes in the computer are intellectual, self-learning processes" (Wisskirchen, Biacabe, Bormann, Muntz, Niehaus, Jiménez Soler & von Brauchitsch 2017, 10). Weak AI is labeled as Artificial Narrow Intelligence (ANI) while strong AI is further distinguished between Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI).

The emergence of robotics technology is developing much quicker than previously thought. Robots are anticipated to soon be as ubiquitous as computers are today (Meghdari & Alemi, 2018). Society has long been concerned with the impact of robotics technology from nearly a century ago, when the word *"Robot" was devised for the first time* (Čapek 1921; Meghdari & Alemi, 2018). The EU Committee on Legal Affairs (2016, p. 4) holds

that "[U]ltimately there is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also to its capacity to be in charge of its own destiny and to ensure the survival of the species." AI mimicking human intellect could soon surpass humans intellectually but also holistically breaking the barrier of human controlled-automization (Schuller 2017; Themistoklis, 2018). Modern literature about robots features cautionary accounts about insufficient programming, evolving behavior, errors, and other issues that make robots unpredictable and potentially risky or dangerous (Asimov, 1942/1950, 1978, 1985; Meghdari & Alemi 2018). "Observe, orient, decide, act" will therefore become essential in the eye of machine learning autonomy and AI forming a new domain of intellectual entities (Armstrong & Sotala 2012, 52; Copeland 2000; Galeon & Reedy 2017; Marra & McNeil 2013). The uncertainty surrounding AI development and self-learning capabilities give rise to the need for guarding AI and an extension of the current legal system to cope with AI (Themistoklis 2018).

With the advancement of technology, social robots have found broader applications in the private and public sectors, such as educational and cultural affairs, games and entertainment, clinical and rehabilitation, nursing of children and/or elderly, search and rescue operations (Meghdari, Alemi, Shariati & Zakipour 2018). For example, social robots such as ASIMO, Nao, iCub, ARASH, and RASA have been developed for "Edutainment" or "education-entertainment" purposes. They aid the study of cognition (both human and artificial), motion, and other areas related to the advancement of robotics serving our society (Meghdari & Alemi 2018). In addition, a few medical and healthcare toy-like robots, such as PARO, which looks like a baby seal, or ARASH, which is a humanoid, have been designed for therapeutic purposes such as reducing distress, stimulating cognitive activity, teaching specific subjects, and improving socialization (Meghdari, Shariati, Alemi & Vossoughi, 2018). Similarly, Sharif University of Technology's socially assistive robot RASA has been developed to help coach and teach Persian Sign-Language to Iranian deaf children (Meghdari, Alemi, Zakipour & Kashanian, 2018). Personal care and companion robots are increasingly

being used to care for the elderly and children, such as RI-MAN, PaPeRo, and CareBot (Meghdari & Alemi,2018). In recent years, robotics technology has extended its applications from factories to more general-purpose practices in society – for instance, such as the use of robots in clinical and rehabilitation, nursing and elderly care, search and rescue operations (Meghdari & Alemi 2018). Social robots have become clinical and educational assistants for social interventions, treatment, and education such as language trainings but also assistance with children with disabilities like autism, down syndrome, cancer distress, hearing impairment, etc. (Meghdari et al., 2018). Initial investigations clearly indicate that social robots can play a positive role in the improvement of children's social performance, reduction of distress during treatments, and enhancing their learning abilities (Meghdari & Alemi, 2018). Surprisingly, although not too hard to imagine, relationships of a more intimate nature have not quite been satisfied by robots yet (Meghdari et al. 2018; Veruggio 2005).

2.2 AI-Evolution

The human perception of and interaction with robot machines with a higher quality physical appearance differs from interaction with a computer, cell phone, or other smart devices (Meghdari & Alemi 2018). For robotics technology to be successful in a human-driven environment, robots do not only need to meet a level of strength, robustness, physical skills, and improved cognitive ability based on intelligence but should also fulfill a social impetus and ethical conscientiousness. The design and construction of social robots faces many challenges, one of the most important is to build robots that can comply with the needs and expectations of the human mind with cognitive capabilities coupled with social warmth (Meghdari & Alemi, 2018). While we have *Social-Cognitive Robotics (SCR)* as a transdisciplinary area of research and a basis for the human-centered design of technology-oriented systems to improve human knowledge functions, judgements and decision making, collaborations, and learning; hardly any information exists on socio-evolutionary comparisons (Meghdari & Alemi 2018). Social-cognitive robotics has been evolving and verified through a series of projects to develop advanced and modern technology-based systems to support learnings and

knowledge functions, and is beginning to play an effective role in societies across the globe (Meghdari & Alemi, 2018). SCR or *Socio-Cognitive Robotics is the interdisciplinary study and application of robots that are able to teach, learn and reason about how to behave in a complex world* (Meghdari & Alemi 2018). Social robotics technology promises a many benefits but also challenges that society must be ready to confront with legal means and ethical imperatives.

2.3 Roboethics

Ethics describes moral principles that govern a person's or group's behavior. Roboethics describes the ethics and morals of robotics, the science of robots. Roboethics therefore captures the integration of ethics into AI and algorithms. This field recently gained considerable attention among humanities and robotics engineers who draw on insights from computer science, artificial intelligence, mechanics, physics, math, electronics, cybernetics, automation and control (Meghdari & Alemi 2018).

What specifies the emergence of socio-cognitive robotics is that humanity is at the threshold of replicating an intelligent and autonomous agent (Meghdari & Alemi,2018). In order to enhance the ability of social robots to successfully operate in humane ways, roles and environments, they are currently upgraded to a new level of physical skills and cognitive capabilities that embrace core social concepts (Meghdari et al. 2018). Robotics thereby unifies two cultures, in which complex concepts – like learning, perception, decision-making, freedom, judgement, emotions, etc. – may not have the same semantic meaning for humans and machines (Meghdari & Alemi 2018).

In the design and construction of social robots, the consideration of ethical concerns has therefore leveraged into an imperative (Lin, Abney & Bekey 2012). Human-robot (a machine with a higher physical and social ability) interactions, are somewhat different compared to other types of human-machine interactions (i.e. with a computer, cell phone, or other smart device) (Meghdari & Alemi 2018; Saffari, Meghdari, Vazirnezhad & Alemi, 2015). It is therefore essential for researchers, scholars, and users to clearly identify, understand, and consider these differences and ethical challenges so that they can benefit from and noone gets harmed by the assistance of social

robots as a powerful tool in providing modern and quality services to society (Meghdari & Alemi 2018; Taheri, Meghdari, Alemi & Pouretemad 2018).

Robots and algorithms now taking over human decision-making tasks and entering the workforce but also encroaching our private lives, currently challenges legal systems around the globe (Themistoklis 2018). The attribution of human legal codes to AI is one of the most groundbreaking contemporay legal and judicial innovations. Until now legal personhood has only been attached directly or indirectly to human entities (Dowell, 2018). The detachment of legal personhood from human being now remains somewhat of a paradox causing an extent of "fuzziness" of the concept of personhood (Barrat 2013; Solum 1992, 1285). As AI gets bestowed with quasi-human rights, defining factors of human personhood will need to be adjusted (Dowell 2018). Human concepts, such as morality, ownership, profitability and viability will have different meaning for AI. The need for redefining AIE has therefore reached unprecedented momentum.

As a predicted trend, the co-existence of AI with the human species is believed to change the fundamental concepts of social, political and legal systems. AI has already produces legal creations and will do so even more in the near future, through its developing autonomy. In addition, the technology leading to AGI and ASI is already present, posing moral and legal dilemmas about who should control it and under what terms (Themistoklis 2018). The emergence of AGI and ASI will necessitate the attribution of some extent and of some type of legal personhood, bearing rights and obligations. AI will not be most probably an exact replication of human intellect behavior (Themistoklis 2018). "[U]ltimately, robots' autonomy raises the question of their nature in the light of the existing legal categories – of whether they should be regarded as natural persons, legal persons, animals or objects – or whether a new category should be created, with its own specific features and implications as regards the attribution of rights and duties" (Committee on Legal Affairs 2016, p. 5). Behavioral economists add the question whether AI and robots should be created to ressemble human beings' decision making with fast thinking and fallible choices or rather be targeted at perfect rationality and slow thinking (Kahneman, 2011). General conscious is strived for so that AI possesses consciousness, which it can evolve and

enhance on the basis of its own critical reflection and assessment of external factors (Themistoklis 2018). A lower level of autonomy exists if an entity can demonstrate such consciousness at a narrow field or can self-evolve and self-adapt to external influences, thus reaching decisions "of its own," without being conscious of its intelligence as such (Themistoklis 2018). As AI emerges as new types of intellect capacities coupled with human-like emotional features, they are attributed a legal personhood in order to ensure to be comprehended correctly and to avoid unfair treatment, towards humans as well (Themistoklis, 2018). Artificial entities are currently gaining human or quasi-human status in the Western and Arab worlds in forming an intellectual autonomy of the entity (MacDonald, 2016). For instance, in Saudi Arabia the first female robot got a citizenship in 2017 and the robot appears to have more rights than a human female in Saudi Arabia (Stone 2017). With the rise of AI persons, their eternal life poses ethical challenges in light of overpopulation and evolutionary perfection, which could crowd out human fallibility if determining merit-based eternal life. These critical questions will be captured in the following.

2.4 Eternal life

While there is currently cutting-edge writing about the potential emergence of an AI personhood as well as concern over the merge of AI with cyberspace that might lead to the breach of the relationship between legal personhood and nation state sovereignty and a nomenclature is emerging on legal characterizations of different levels of AI development; hardly any information exists about the eternal living of AI (Hildebrandt 2013). From the theoretical standpoint, the eternal longeavity of AI contradicts the fundamental concept of fairness in death, as a general condition for all. From the practical standpoint, the international community is currently urged to think on the basis of global commons in terms of AI and AI eternal life potentials contributing to overpopulation. Thereby global commons theories may be tabbed on, which primarily offer guidance for a regulatory framework, which establishes control "…for the benefit of all nations" and refer to space constraints (Clancy 1998; Tsagourias 2015).

Regarding limited space, longeavity and eternal life appears problematic. Humankind may face tough decisions whether or not to have AI proceed and what kind of developments to flourish and what to extinct. In what cases should we consider to switch off AI? In 1950, Isaac Asimov introduced the idea robot to (1) not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot obeying the orders given it by human beings except where such orders conflict with the first law. (3) A robot must protect its own existence as long as such protection does not conflict with the first or second law. So in the cases of overpopulation and harm emerging from AI, algorithms and robots can be considered to be switched off. But when to stop AI?

Another killing market mechanism may be natural market selection via price mechanisms and the falling rate of profit. Regarding prices, natural supply and demand mechanisms will always favor innovation with a higher price and following supply of goods lead to a price drop. The falling rate of profit is one of the major underlying features of business cycles, long-term booms and downturns (Brenner, 2003, 2006a, b). Capitalism is thereby described as competitive battle for innovation and reaping benefit from first-market introductions. Once followers enter the market, profit declines, leading eventually to market actors seeking novel ways to innovate in order to regain a competitive market advantage and higher rates of profit. Thereby industries and innovations fade and die off. Such a natural market evolution is also likely to occur with AI innovations, which will determine which AI traits will remain and which ones will fade off.

Apart from soft market mechanisms that may lead to AI evolution, what are the cases when AI should be shut down or switched off or – in the case if AI personhood – be killed?

2.5 Death

Errors and Safety: When errors occur and general safety is at stake. The main and leading concern about any new and emerging technology is to be safe and error free (Meghdari & Alemi, 2018). Therefore, sufficient and numerus tests on health and safety must be performed by developers and/or well-known independent sources before rolling out any technology

onto the marketplace and society (Meghdari & Alemi, 2018). In robotics, the safety issue mainly centers around software and/or hardware designs (Meghdari & Alemi, 2018). Even a tiny software flaw or a manufacturing defect in an intelligent machine, like a smart car or a social robot, could lead to fatal results (Meghdari & Alemi 2018). When these deviations occur and especially when they are harmful to the human community but also to other AI species, the faulty AI should be terminated. With regard to the risk of robotic malfunctions and errors, product legal responsibility laws are mostly untested in robotics (Meghdari & Alemi 2018). A usual way to minimize the risk of damage from social robots is to program them to obey predefined regulations or follow a code-of-ethics (Meghdari & Alemi 2018). Ethical codes for robotics are currently needed and should become formed as a natural behavioral law to then be defined and codified as law. Laws but also an ethical understanding to terminate AI, algorithms and robots in case of impairment and harm are needed.

Morals, Ethics, and the Law: As social robots become more intelligent and autonomous and exhibit enough of the features that typically define an individual person, it may be conceivable to assign them responsibility and use them in social, educational, and therapeutic settings (Meghdari & Alemi, 2018). In the currently ongoing research on the integration of computers and robotics with biological corpse it is found that a cognizant human brain (and its physical body) apparently has *human-rights*; hence, replacing parts of the brain with artificial ones, while not harming its function, preserves those rights (Meghdari & Alemi 2018; Warwick & Shah 2014). Also, consider a handicapped person featuring an electronic robot arm that commits a crime. It becomes obvious that half-robot-human beings should be considered as human and robots as quasi-human beings. Meghdari & Alemi (2018) speculate that at some point in the future, we may face a situation in which more than half of the brain or body is artificial, making the organism more robotic than human, which consolidates the need of special *robot-rights* and attributing (quasi)-human rights onto robots. When considering robots as quasi-human beings, their termination appears legally questionable and ethically challenging, requiring to revisit laws as legitimation to kill a likewise species as well as ethical consensus on the virtue of killing.

The legal argumentation may draw on justifiable homicide as outlined in criminal law cases – such as prevention of greater harm to innocents during an imminent threat to life or well-being in self-defense. According to the United Nations Universal Declaration of Human Rights, Article 3 states that everyone has the right to life, liberty and security of person and most nations' policy allows for some degree of leniency for self-defense, which reduces charges.** Potentially excusing conditions common to most jurisdictions include wartime, when the person's death is influcted by the effect of a lawful arrest or prevention of lawfully detained person's escape, quelling riot or insurrection, when the use of force is„no more than asolutely necessary." Some countries deem it lawful for a citizen to resort to violence to protect valuable property and there is the "heat of the moment" defense argument, in which the defendant deemed to have lost control through provocation. Doctrine of necessity allows, for example, a surgeon to separate conjointed twins and killing the weaker twin to allow the stronger twin to survive. While fetuses are considered as unborn children in the US, the right to an abortion was upheld in the US legal system as exeption from prosection (*Roe v. Wade, 1973*). Several countries, such as the Netherlands, Belgium, Switzerland, Japan, and the U.S. states of Oregon and Washington, allow both active and passive euthanasia by law, if justified.

Where the person concerned is to be arrested for an offense referred to in Schedule 1 or is to be arrested on the ground of having committed such an offense, and the person authorized under this Act to arrest or to assist in arresting him cannot arrest him or prevent him from fleeing by other means than killing him, the killing shall be deemed to be justifiable homicide.

If any arrestor attempts to arrest a suspect and the suspect resists the attempt, or flees, or resists the attempt and flees, when it is clear that an attempt to arrest him or her is being made, and the suspect cannot be arrested without the use of force, the arrestor may, in order to effect the arrest, use such force as may be reasonably necessary and proportional in the circumstances to overcome resistance or to prevent the suspect from fleeing: Provided that the arrestor is justified in terms of this section in using deadly force that is intended or is likely to cause death or grievous bodily harm to a

** http://www.un.org/en/universal-declaration-human-rights/

suspect, only if he or she believes on reasonable grounds (§7 Judicial Matters Second Amendment Act 122 of 1998).

In light of overpopulation and harmful behavior of AI, switching off artificial life, which is currently be granted quasi-human status, will need to be argued legally and supported ethically. Killing in terms of the death penality is justified legally in the 5th (and the 14th) amendment that states "no person shall be deprived of life, liberty, or property without due process of law," while the 8th amendment prohibits "cruel and unusual punishment."

Killing in terms of harmful behavior of AI can be grounded on similar legal reasons to ensure that no AI harms the collective. Overpopulation claims leading to the need to take AI partially off the grid more lead to philosophical sources that argue for individual's free will to choose to live or die (Critchley, 2015; Critchley & Hume, 2016). Suicide has been tabooed for most part of history and propagated to be a religious sin. Yet the human gift of reflection and search for meaning in life or death could leverage into an asset in the AI evolution in the decades to come. We could argue that similar to critique on those who proclaim loudly against suicide and claim that the act of taking one's own life is irresponsible and selfish, even shameful and cowardly, that people must stay alive whatever the cost (Critchley 2015; Critchley & Hume 2016); there will be virtue in the killing AI. Suicide understood as neither a legal nor moral offence but as right to life or death bestowed upon human beings in their self-conscious reflection may be extended as a virture of killing in the artificial age, when human beings will have to decide what AI should stay alive and what AI be taken off the grid. Human will thereby become the rulers of the forthcoming AI evolution.

The virture of killing could also be grounded on Viktor Mayer-Schönbergers "right to be forgotten," which ensures data privacy through automated deletion of contents after a certain period and grants individuals rights to have their data been destroyed (Puaschunder 2018a, forthcoming). However, the implementation of this right is still in infancy and hindered by questions of what court is responsible for an as such claim. As a legal subsumption, we may speculate that individuals may be granted a ‚right to terminate' and can order for robots to be switched off if causing harm to them. As the ‚right to be forgotten' law can be overruled by concern for public safety,

this may also appy to the right to terminate. Thereby it deserves mentioning that safety differs around the world and also expected safety standards.

2.6 AI-Evolutionary pressure turning against human

The predicted AI-Evolution (AIE) is grounded on evolution as the change in heritable characteristics of biological populations over successive generations. As for human evolution, these characteristics are the expressions of genes that are passed on from parent to offspring during reproduction. Different characteristics tend to exist within any given population as a result of mutations, genetic recombination and other sources of genetic variation. Evolution occurs when evolutionary process such as natural selection (including sexual partner selection) and genetic drift act on these variations, resulting in certain characteristics becoming more common or rare within a population. This process has given rise to biodiversity at every level of biological organisation including the levels of species, individual organisms and molecules. Evolution by natural selection defines the following facts about living organisms: Traits vary among individuals with respect to their morphology, physiology and behavior (phenotypic variation). Different traits confer different rates of survival and reproduction (differential fitness). Traits are passed from generation to generation (heritability of fitness). Thus, in successive generations members of a population are more likely to be replaced by the progenies of parents with favorable characteristics that have enabled them to survive and reproduce in their respective environments.

AIE now refers to the human process of selecting what AI should survive or be killed by being taken off the grid forming heritable characteristics of blockchain-like created populations of robots and AI. Like genes being passed on from parents through natural mate selection, decision makers will divert favorable traits from unfavorable. Mutations may occur in decision making errors innate in human beings as described by behavioral economics (Puaschunder 2017a). AI traits will be varying in their survival rate. Favorable characteristics will have a higher likelihood to survive. But what will count as favorable will be determined by human and therefore add a social touch to future AI to come. However, the critical problem appears that robots will outperform human beings and could turn around evolutionary

pressures towards the eradication of the fallible species of human. In the creation of AI, stereotypes should be eradicated and a social class division avoided (Puaschunder b, c, d).

3. Discussion

The growing number of AI and robotics researchers are demanded to create a framework on AI ethics building on the benefits of humanities, philosophy, sociology, and social neuroscience expertise and research. Likewise, growing trends of mutual collaboration among scholars in the field of human sciences, linguistics, and psychology with the robotics scientists are producing quite noticeable valuable results (Meghdari & Alemi 2018). Future studies should target at presenting an overview of the novel and multidisciplinary area of socio-cognitive robotics, and further explore the possible ethical challenges of emerging technologies on education, culture, entertainment, gaming, nursing, and therapy. Unraveling ethical features based on our past and present research experiences in a variety of areas will aid designing safe AI and social robots.

In its entirety, this article was the first introduction of AI ethics opening up many challenging questions. For instance, what ethical code should we apply for controlling robots' actions? How can we program a switch to turn off AI in case of unlawful action and harm to people but also how to draw the boundery condition to ethical infringements? This is specifically important if humankind starts placing social robots in positions of authority, such as police, security guards, teachers, or any other government roles or offices, in which humans would be expected to follow them.

In the further discussion of the topic, research should analyze the effects of robotics blending into our societies with direct applications in fields where the potential complications are more significant and apparent (Meghdari & Alemi, 2018). Important areas of scrutiny should be human rights/dignity, equality and justice, benefits and damage, cultural diversity and pluralism, religious variety, non-discriminating, independence and individual accountability, privacy and confidentiality, unity and collaboration,

social responsibility, benefits sharing and environmental obligations as well as intergenerational equity considerations (Meghdari & Alemi 2018).

4. Conclusion

The days of AI being a futuristic concept are over. AI is now. Social and cognitive robotics is rapidly becoming one of the leading fields of science and technology involving a deep level of human-machine interaction (Meghdari & Alemi 2018). he world will soon be populated with human and machines alike that will coexist. The clear advantage of AI is the longeavity. In light of overpopulation fear, we need mechanisms to determine how to decide over what is worth living forever and what should be taken off society. Ethics may come into this predicted AI-evolution. One may conclude that roboethics entails the ethics of handling and application of robots (Meghdari & Alemi 2018).

It is predicted that society is expected to fall into two extremes of a dichotomy between rationality (represented by AI) and humanness (represented by human beings). Hereby the question arises what is it that makes human humane? In the age of artificial intelligence and automated control, humanness is key to future success. Behavioral human decision making insights and evolutionary economics can already today predict what makes human humane and how human decision making is unique to set us apart from artificial intelligence rationality. Future research in these domains promise to hold novel insights for future success factors for human resource management but also invaluable contributions for artificial intelligence ethics (Puaschunder 2018b).

Overall this paper was meant as first step towards a nomenclature of deciding on the future evolution grounded in the virtue of living and killing to motivate different viewpoints on the issue by cultural, religious, and ethical scholars. The article plays an important role in the evolution of an AI and human mixed society in order to ground stability and social harmony into the newly emerging system. Depicting ethical imperatives around the life and death of machines being considered as quasi-human beings during this unprecedented time of societal change and regulatory

reform holds invaluable historic opportunities for global governance policy makers to snapshot the potential but also save from the likely downfalls of a robo-human mixed society.

The results are targeted at guiding a successful introduction to AI and lower systemic downfalls with attention to the changes implied in the wake of the ongoing artificial intelligence revolution. Market and societal policy recommendations for global governance experts on how to strengthen society but also overcome unknown emergent risks within globalized markets and bestow market actors with key qualifications in a digitalized world are endeavored alongside scientific publications and stakeholder engagement.

In the international compound, having parts of the world being AI-driven and others being human capital grounded is prospected to increase the international development divide in the years to come. While in the AI-hubs human will be incentivized become more creative and humane while AI performs all rational tasks to a maximum productivity, other parts of the world will naturally fall back as for being stuck in spending human capital time on machine-outsourceable tasks and not honing humane skills, which are not replicable by machines. All these endeavors promise challenging ethical, social, and economic controversies.

It constitutes a matter of the present as well, given that the technology leading to autonomous GAI and SAI is present and evolving challenging contemporary questions for humankind. The regulation of the current technological advancement needs an integration of multi-faceted problem solving approaches. On the basis of these assumptions, it is suggested that the regulatory framework of terminating AI should be centered around a global commons theory and because of its unique nature needs to borrow elements of normative frameworks of different fields other than law, such as philosophy and urban planning. In addition, the framework of global commons could establish a transparent framework for the regulation of technological advances, leading to the unique situation of the emergence of non-human, autonomous, intellect beings, bestowed with legal personhood and ready to be killed.

References

Alemi, M., Meghdari, A. & Saffari, E. 2017. "RoMa: A hi-tech robotic mannequin for the fashion industry." *Lecture Notes in Computer Science (LNCS): Social Robotics*, 10652, 209-219.

Armstrong, St. & Sotala, K. 2012. "How we're predicting AI – or failing to." In: J. Romportl (Ed.), *Beyond AI: Artificial Dreams*, 52. Pilsen: University of West Bohemia.

Asimov, I. 1942/1950. *I, Robot*. New York: Bantam Dell.

Asimov, I. 1978. My own view. In: R. Holdstock (Eds.), *The Encyclopedia of Science Fiction*, N.Y.: St. Martin's Press.

Asimov, I. 1985. *Robots and empire*. New York: Doubleday.

Barrat, J. 2013. *Our final invention: Artificial Intelligence and the end of the human era*. New York: St. Martin's Press.

Beerbaum, D. & Puaschunder, J.M. 2018. "A behavioral economics approach to digitalization: The case of a principles-based taxonomy." In *Proceedings of the 9th International RAIS Conference on Social Sciences and Humanities* organized by Research Association for Interdisciplinary Studies (RAIS) at The Erdman Center at Princeton University, Princeton, New Jersey, USA, August 22-23, 2018.

Brenner, R. 2002. "American economic revival." In R. Brenner, *The Boom and the Bubble: The US in the World Economy*. New York: Verso.

Brenner, R. 2006a. "The puzzle of the long downturn." In R. Brenner, *The Economics of Global Turbulence: The Advanced Capitalist Economies from Long Boom to Long Downturn, 1945- 2005*. New York: Verso.

Brenner, R. 2006b. "From boom to downturn, In R. Brenner." *The Economics of Global Turbulence: The Advanced Capitalist Economies from Long Boom to Long Downturn, 1945-2005*. New York: Verso.

Cellan-Jones, R. 2014. "Stephen Hawking warns artificial intelligence could end mankind." *BBC News*, 2 December. www.bbc.com/news/technology-30290540.

Căpek, K. 1921. *Rossum's universal robots*. New York: Penguin.

Clancy, E. 1998. "The tragedy of the global commons." *Indiana Journal of Global Legal Studies* 5, 2, 601-619.

Committee on Legal Affairs. 2016. *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*. May 31. 2015/2103(INL)

Copeland, J. 2000. What is Artificial Intelligence? *AlanTuring.net*, May. www.alanturing.net/turing_archive/pages/ Reference%20Articles/ what_is_AI/What%20is%20AI02.html.

Critchley, S. 2015. *Suicide.* London: Fitzcarraldo Editions.

Critchley, S. & Hume, D. 2016. *Notes on suicide*. London: Fitzcarraldo Editions.

Dowell, R. 2018)."Fundamental protections for non-biological intelligences or: How we learn to stop worrying and love our Robot Brethren." *Minnesota Journal of Law, Science & Technology*, 19, 1, 305-336.

EU Committee on Legal Affairs. 2016. *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*, May 31.

Galeon, D. & Reedy, Ch. 2017. "Kurzweil claims that the singularity will happen by 2045." *Futurism*, October 5, futurism.com/kurzweil-claims-that-the-singularity-will-happen-by-2045/

Hayes, A. 2018. *The active construction of passive investors: Toward Robo economicus*." Working paper, University of Wisconsin-Madison: Department of Sociology.

Hildebrandt, M. 2013. "Extraterritorial jurisdiction to enforce in cyberspace? Bodin, Schmitt, Grotius in cyberspace?" *Toronto Law Journal* 63, 196-224.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kowert, W. 2017. The foreseeability of human-artificial intelligence interactions. *Texas Law Review* 96, 181-204.

Larson, D.A. 2010. "Artificial Intelligence: Robots, avatars, and the demise of the human mediator." *Ohio State Journal on Dispute Resolution* 25, 105-164.

Laton, D. 2016. "Manhattan_Project.Exe: A nuclear option for the digital age." *Catholic University Journal of Law & Technology* 25, 4, 94-153.

Lin, P., Abney, K. & Bekey, G.A. 2012. *Robot ethics: The ethical and social implications of Robotics*. London, England: The MIT Press.

MacDonald, F. 2016. Harvard scientists think they've pinpointed the physical source of consciousness. *Science Alert*, June 23. http://www.sciencealert.com/harvard-scientists-think-they-ve-pinpointed-the-neural-source-of-consciousness.

Marra, W. & McNeil, S. 2013. "Understanding "the loop": Regulating the next generation of war machines." *Harvard Journal of Law & Public Policy*, 36, 1139-1187.

Meghdari, A. & Alemi, M. 2018. "Recent advances in social & cognitive robotics and imminent ethical challenges." In *Proceedings of the 10th International RAIS Conference on Social Sciences and Humanities* organized by Research Association for Interdisciplinary Studies (RAIS) at The Erdman Center at Princeton University, Princeton, New Jersey, United States. Cambridge, MA: The Scientific Press.

Meghdari, A., Alemi, M., Zakipour, M. & Kashanian, S.A. 2018. "Design and realization of a sign language educational humanoid robot." *Journal of Intelligent & Robotic Systems*, 1-15, Springer, 2018.

Meghdari, A., Shariati, A., Alemi, M. & Vossoughi, G.R. 2018. "Arash: A social robot buddy to support children with cancer in a hospital environment." *Journal of Engineering in Medicine 232* (6): 605-618.

Noyes, K. 2016. "5 things you need to know about A.I.: Cognitive, neural and deep, oh my!" *Computerworld*, March 3. Retrieved at www.computerworld.com/article/3040563/enterprise-applications/5-things-you-need-toknow-about-ai-cognitive-neural-anddeep-oh-my.html.

Puaschunder, J.M. 2017a. "Nugitize me! A behavioral finance approach to minimize losses and maximize profits from heuristics and biases." *International Journal of Management Excellence*, *10*, 2, 1241-1256.

Puaschunder, J.M. 2017b. "Nudging in the digital big data era." *European Journal of Economics, Law and Politics 4*, 4, 18-23.

Puaschunder, J.M. 2017c. "Nudgital: Critique of Behavioral Political Economy." *Archives of Business Research* 5 (9): 54-76.

Puaschunder, J.M. 2017d. "The nudging divide in the digital big data era." *International Journal of Research in Business, Economics and Management 4*(11):12, 49-53.

Puaschunder, J.M. (2018a). A utility theory of privacy and information sharing. Social Science Research Network working paper, Retrievable at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3197744

Puaschunder, J.M. (2018b). Artificial Intelligence Ethik. Social Science Research Network paper. Retrievable at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3137926

Puaschunder, J.M. (forthcoming). Towards a utility theory of privacy and information sharing and the introduction of hyper-hyperbolic disounting in the digital big data age. *Encyclopedia of Information Science and Technology*.

Russell, St. & Norvig, P. 1995. *Artificial intelligence a modern approach*. New Jersey: Simon & Schuster.

Saffari, E., Meghdari, A., Vazirnezhad, B. & Alemi, M. 2015. Ava (A social robot): Design and performance of a robotic hearing apparatus. *LNCS: Social Robotics*, 9388, 440-450, Springer, Oct. 2015.

Schuller, A. 2017. "At the crossroads of control: The intersection of artificial intelligence in autonomous weapon systems with international humanitarian law." *Harvard National Security Journal* 8: 379-425.

Sofge, E. 2015. "Bill Gates fears A.I., but A.I. researchers know better." *Popular Science*. Retrieved at www.popsci.com/ bill-gates-fears-ai-ai-researchers-know-better.

Solum, L. 1992. Legal personhood for artificial intelligences. *North Carolina Law Review*,*70*(4): 1231-1287.

Stone, Zara. 2017. "Everything You Need To Know About Sophia, The World's First Robot Citizen." *Forbes*, https://www.forbes.com/sites/zarastone/2017/11/07/everything-you-need-to-know-about-sophia-the-worlds-first-robot-citizen/#7301bab246fa.

Taheri, A.R., Meghdari, A., Alemi, M., Pouretemad, H.R. 2018. "Human–robot interaction in autism treatment: A case study on three pairs of autistic children as twins, siblings, and classmates." *International Journal of Social Robotics 10*(1): 93-113.

Themistoklis, T. 2018. "Artificial intelligence as global commons and the "international law supremacy" principle." In *Proceedings of the 10th International RAIS Conference on Social Sciences and Humanities* organized by Research Association for Interdisciplinary Studies (RAIS) at The Erdman Center at Princeton University, Princeton, New Jersey, United States. Cambridge, MA: The Scientific Press.

Tsagourias, N. 2015. "The legal status of cyberspace." In N. Tsagourias & R. Buchan (Eds.), *Research Handbook, International Law and Cyberspace*, pp. 13-29. Cheltenham: Edward Elgar Publishing.

United Nations Department of Economic and Social Affairs. 2017. *Will robots and AI cause mass unemployment? Not necessarily, but they do bring other threats*. New York: https://www.un.org/development/desa/en/news/policy/will-robots-and-ai-cause-mass-unemployment-not-necessarily-but-they-do-bring-other-threats.html.

Veruggio, G. 2005. "The birth of roboethics." *ICRA 2005, IEEE Int. Conference on Robotics and Automation: Workshop on Robo-Ethics*, Barcelona, April 18, 2005.

Warwick, K. & Shah, H. 2014. "How good robots will enhance human life." In K.Tchoń & W.W. Gasparski (Eds), *Treatise on Good Robots Edition: Praxiology: The International Annual of Practical Philosophy and Methodology* Vol. 21, Book Chapter, Transaction Publishers, USA Editors: January 2014.

Wisskirchen, G., Biacabe, B.T., Bormann, U., Muntz, A., Niehaus, G., Jiménez Soler, G. & von Brauchitsch, B. 2017. *Artificial Intelligence and robotics and their impact on the workplace*. London: IBA Global Employment Institute.

Scientia Moralitas International Journal of Multidisciplinary Research
ISSN 2472-5331 (Print) | ISSN 2472-5358 (Online) | Vol. 4, No. 1, 2019

The Need for an International Treaty for AI from the Perspective of Human Rights

Themistoklis Tzimas, PhD
University of Macedonia, Thessaloniki, Greece
themis.tzimas@gmail.com

ABSTRACT: The article analyzes the role of human rights in relation to Artificial Intelligence. The main goal is to identify how human rights can contribute into a new international treaty, attempting to regulate the advances and the functions of AI, both at the present, narrow field, as well as at the level of general or super intelligence in the future. In order to do so, the article examines issues which are related to the ontology of AI, which determine the transformation of social and subsequently of legal relations too. In such a framework, the impact of human rights is presented.
KEYWORDS: Artificial Intelligence, autonomy, human rights, international law

Introduction

The present and future technological, as well as social, economic and political developments are already and will be further defined by the rise of Artificial Intelligence.(Ben-Ari, Frish, Lazovski, Eldan & Greenbaum 2017, 10). Several developments – i.e. the emergence of the so- called"fourth industrial revolution" or issues related to intellectual property and patents, military operations, arts, education, medicine, governance, social policy making, finance, environment and the equivalent fields of law being some of them- indicate such a defining role.

AI explosive expansion has raised both concerns (Sofge 2015) and expectations because of historically novel and unique issues (Larson 2010, 106; Kowert 2017, 181-83).

At the core of these issues lays the unique ontology of AI, which is built on the growing and expanding autonomy of AI entities, which both complicates the relationship of AI and humans from the perspective of the latter and raises the potential for a new type of legal personhood, that of AI. In this sense, human rights become critical in terms both of a potential AI legal personhood and of humans' protection. This is the framework of the present examination.

In order to examine the role of human rights, the paper first analyses the ontology of AI. It then examines the relevance of human rights, applying them to the ontologies of AI.

1. The AI "ontology"

AI ontology is surrounded by ambiguity at a significant extent. "[I]n spite of what I regard as AI's significant achievements . . . the not so well-kept secret is that AI is internally in a paradigmatic mess" Chandrasekaran comments. (Chandrasekaran 1990, 14). The definition of AI is debatable too (Russell & Norvig 2013, 2). It has been defined as "a broad set of methods, algorithms, and technologies that make software 'smart' in a way that may seem human-like to an outside observer" (Noyes 2016) A slightly different definition describes AI as "Machines that are capable of performing tasks that, if performed by a human, would be said to require intelligence" (Scheree 2016, 363-64).

AI definitions include the elements of "consciousness, self-awareness, language use, the ability to learn, the ability to abstract, the ability to adapt, and the ability to reason" (Scheree 2016, 363-64) of goal orientation and of the rational agent (Russell & Norvig 2010, 2-3). The focus of most definitions lays in the "human- like" intelligence of machines, although that can be partially deceiving, as an entity mimicking human intelligence does not necessarily "understand" or share the patterns of human intellect (Laton 2016, 94).

AI is distinguished between weak AI, where "the computer is merely an instrument for investigating cognitive processes" and strong AI, where "[t[he processes in the computer are intellectual, self-learning processes"(Wisskirchen 2017, 10). Weak AI is labeled as Artificial Narrow Intelligence-ANI- while strong AI is further distinguished between Artificial General Intelligence –AGI– and Artificial Super Intelligence – ASI (Urban 2015). It must be noted however that ANI has already surpassed the direct control from the programmer too.

Therefore, the learning procedure and autonomy already exist having surpassed the automation phase; however until now they apply only in specific areas, unlike humans who possess general intelligence. Although AI has already"outsmarted" humans in certain, narrow areas and tasks, it cannot –yet- compete with humans, in terms of adaptable and general intelligence.

AGI will be consisted of the "type of adaptable intellect found in humans, a flexible form of intelligence capable of learning how to carry out vastly different tasks… based on its accumulated experience"(Heath 2018) enabling it to choose by itself, where and how to apply its intelligence. The "when" of AGI is debatable, although most analysts agree that within this century it will happen (Tal 2018). Super intelligence refers to the exceeding of human intelligence in the sense of"…an intellect that is much smarter than the best human brains in practically every field..."(Bostrom 1998).

While the time of the achievement of super intelligence remains at stake, its achievability is foreseen with some certainty. As an article co-authored by Stephen Hawking, Max Tegmark, Stuart Russell, and Frank Wilczek foresaw that: "…there is no physical law precluding particles from being organized in ways that perform even more advanced computations than the arrangements of particles in human brains" (Hawking et al. 2014).

The main idea is that since human brain performs computation, a different, non- biological computational entity could perform like the human brain and eventually out-perform it (Snyder- Beattie & D. Dewey 2014). At the core of AI development lays the intellectual autonomy of the entity, in combination with developments such as big data, better algorithms and improved hardware (MacDonald 2016). Intellect autonomy is built on "machine- learning", comprised of a performance and of a learning element.

The first one "senses the environment", while the latter, employs feedback from the system and amends the performance element (Marra & S. K. McNeil 2013, 1145).

Machine learning thus resembles more to "coaching" than programming (Tanz 2016; Scherer 2016, 33) and also to human learning procedure (Schuller 2017, 404). It can be also described through the cumulative contribution of three abilities: to compute information, to learn and to reason (Khoury 2017, 640).

Machine learning is already giving way–at least up to some extent–to neural networks and deep learning. Neural networks are inspired by human brain and the synapses between neurons, which function at different layers, through which, massive data run, in order to train the system. An AI neural network is a "biologically inspired computational model that is patterned after the network of neurons present in the human brain", modeling "the input-output relationship" (Nvidia 2019). Neural networks sustain and enhance machine learning, promoting and accelerating AGI.

In the framework of such procedure, AI entities need to include various components, such as logic- "as a tool of analysis, as a basis for knowledge representation, and as a programming language"(Thomason 2003) —creativity—combined with skills such as problem solving, pattern recognition, classification, learning, induction, deduction, building analogies, optimization, surviving in an environment and language processing (Hutter 2010, 125-126, 231) —communicative capacities, external knowledge, "cognitive autonomy" —in the sense of working "independently without human intervention beyond defining goals" - intuition and strategic thinking (Camett & Heinz, 2006; Suchman and J. Weber 2016, 39-40).

Machine learning and neural networks have already surpassed "rules-based programming",(Pyle & C. San Jose 2015) providing AI the capacity to function autonomously from the human programmer, surpass by far human intelligence –*currently*– in narrow, pre-determined areas, evolve and even re-programme itself. Of course, AI has not yet achieved general intelligence and is still indicating these exceptional capacities, in a "protected" environment.

Much higher autonomy will take place when AI entities will be endowed with self- awareness, in the sense of being aware of their own

existence and of placing themselves in the broader world, with –as mentioned above- adaptable intelligence which may lead to their choices not only in terms of means but also in terms of goals (Chong 2015; Schkolne 2018). Such conception of self- awareness implies a unity of subjective, mental activities, such as imaginative thinking, self- decision, creativity, self- representation and self- discovery, sentience, wakefulness, all of which tend to re- inventing one's own presence in the world. These elements describe aspects of consciousness (Herbert 1985, 249) with the latter comprehended as "…self-reflective… [as] the perception of perception, and the awareness of awareness"(Smith 1998, 281; Tegmark, 2018, 428-30, 431). Essentially, consciousness is condensed in the subjective experience, which also bears with it a certain degree of unpredictability.

Such development however should not be perceived as necessarily leading to intellect- autonomy and function, identical to that of humans. On the contrary, it is likely that the concepts of the "self" and of the surrounding environment may be inherently different for AI (Damasio 1994, 247-248). While it is with AGI and ASI that the fore- mentioned issue becomes emphatically present, it is also present with existing, AI intellect autonomy at relatively narrow fields, which can produce impressively beneficial or destructive consequences, both unpredictable and not traceable or attributable to the initial human programmer (Eden, Steinhart, Pearce & Moor 2012, 28-9; Del Prado 2015; Bostrom 2014, 26-29, 140, 155).

Summing up, the argument is that the developing ontology of AI is condensed in its expanding autonomy which tends towards subjectivity and therefore unpredictability, the extent of which is determined on the basis of intellect capacity, adaptability and generalization, as well as of autonomy. This is why the argument of the present article is that a new framework specifically designed for AI, both in its current and in its potential forms is immanently necessary.

2. A regulatory framework for AI- the role of human rights

On the basis of the above- mentioned ontological elements and of the prospect they bear to fundamentally alter human conducts or even to introduce us into an era of new "beings" and legal subjects, of non- human

orientation, the need for a legal framework, capable of present and future developments.

Until now, there are only mild and primary efforts for the establishment of a legal framework, as well as declaratory documents by private entities. Indicatively, the EU Parliament adopted a resolution about civil law rules on robotics, endorsing Asimov rules for autonomous AI and robotics (European Parliament 2017).

Other powers, such as the US, China and the UK are also working on regulatory frameworks, without having produced though coherent legal frameworks. Private institutions have contributed into the gradual formation of more de- centralized regulatory schemes, which however cannot be substitutes to full- fledged, legal schemes (Triolo P., Kania E., and Webster G., 2018; Black 2001, 103).

The answer to the question about the proper type of legal regulation must be determined on the basis of novelty, of risk and of expansion of AI. The novelty determines the extent of suitability of the existing legal systems; the risk factor, determines the prevalence of hard or soft and de- centralized law approaches; the impact, the main "beneficiaries" of the regulation; It is on the basis of a combined approach to these criteria that we reach the conclusion that novel and adaptable legal systems are required, in the sense of an international treaty so as to avoid fragmented and therefore inadequate responses (Andersen 2018, 55-56).

Existing legal systems can contribute with existing fundamental principles—albeit in some cases with the necessary changes—in order to achieve a three- end goal: preserve the safety and the rights of humans, preserve fairness among humans and when AGI will have been achieved preserve the rights which will be flowing from the potential legal personhood of AI entities. In this sense human rights, as existential rights for humans and for the international community, set the ultimate checks and balances for legal systems and therefore, potentially for the regulation of AI too (Alston 1984, 607).

Human rights can establish a regulatory framework that will be prohibiting and enabling certain AI developments and applications and also they must constitute a positive obligation of programmers, manufacturers

and owners of AI in the sense of "training" of AI systems so that they endorse the overall goals and the specific, human rights.

However, the actual implementation of human- rights' guided and trained AI will have more complexities than it seems: the growing autonomy means that the effectiveness of "training" of AI entities may eventually be proven limited and also we cannot yet foretell how a non- human, intelligent entity will comprehend in its self- development and self- conscious course, human rights. We can try and create "friendly" AI, meaning AI that will share "our" goals and our idea of humanity and of the preserve it. However we can never be absolutely certain that such guarantees will be proven efficient even in ANIand we cannot rely solely on a training procedure without a more general and intervening, regulatory framework, in different stages of AI evolution (Omohundro 2008, 483-92).

Therefore, the prospect of intelligent entities, which may be equally intelligent or superior to us, posing existential danger, could justify a slowing down or even a prohibition of certain technological advances, which lead to AGI and ASI, via a relevant treaty, establishing that AI technology that can be threatening for the superiority of human intelligence and for the goals of the international community will be prohibited (De Garis 2005, 1-2).

Such an approach however—if chosen—has the defect that it solely emphasizes upon the potential risk from AI, being therefore up to some extent, one- sided while AI applications can be double- edged; both beneficial and possibly harmful. In some sense, AI according to analysts can be proven even morally enhancing to humans (Waser 2008).

Therefore what is proposed is the intervention in advance and if needed in "correction" of the four main reasons for unethical behavior: namely "over-riding self-protection (fear); selfishness (greed); unfairness (error) on society's part; or error on the entity's part" (Waser 2008). If the ethical risk can be minimized, a general prohibition of certain AI developments will rather harm than safeguard humanity and human rights too. We need therefore to imagine a more elaborate and complicated legal system, which will be able to provide better guarantees regarding- among other areas of law- the guidance of AI by human rights as well as to capture the potentially beneficial and benevolent impact of AI, without undermining the risks too.

The first principle of such an approach must be that human rights should guide the technological research and the applications of AI, as a positive obligation of manufacturers, programmers and owners of AI to train the latter in line with human rights. Therefore, the flow of big data, the algorithms and software that are used must include human rights as part of machine learning and of the training procedure.

The second principle should refer to the differentiation among the various AI applications- actual or potential- and to technological research leading to them. It cannot be overlooked that there are applications which tend to be more beneficial for humans and for the promotion of human rights, whereas others bear more risks. Depending on the potential risk to human rights–among other things- that they represent they can be divided between low, medium and high risk AI.

Such categorization can be determined on the basis of the goals, as well as of the means and will be leading to policies of further promotion, of partial restriction or of prohibition of certain applications—actual or future—and of technology leading to them, depending on the risk that they pose. There may be several and different policies and measures, such as the control of the type of data provided or the disconnection of certain AI applications from the cyberspace or parts of it.

The third principle of a potential legal regulation, on the basis of human rights, engulfs the most intriguing issue, which is that of the regulation of the potential emergence of AGI and of ASI. Can the path towards such developments be legitimate under human rights imperatives? The answer to the question is pre- legal: if the prevalent assumption is that AGI or and ASI will certainly or likely become hostile towards humans, then human rights impose the obligation to terminate research moving towards this direction, at least "one step" before reaching any of these two levels. Otherwise, we must focus upon these checks and balances, in accordance with human rights so that we keep it non- hostile and beneficial for us, enhancing its benevolent tendency.

In case however AGI and ASI is eventually achieved, human rights will have to adapt given that most likely there will be an international or- to better present it- a global community comprised from human and non-

human being of equal or superior intellect capacity. While human rights may be able to retain their relevance for humans they will stop constituting the fundamental norms of that new, global community.

One last thing that remains to be discussed is how human rights will be related with the potential legal subjectivity of AI, in case the latter is achieved (Lawson 1957, 915; Solum 1992, 1285; Barrat 2013, 39-41; Dowell 2018, 321, 327-29). In this sense, all legal systems are human- centric and take for granted that humans are the dominant and more developed form of being- intellectually speaking- the welfare of who constitutes the main goal. The impetuous development of AI can challenge this, until now, self- obvious fact (Anderson M. & Anderson S. L. 2011, 7-13).

Up to the extent that conscience, reason, self- awareness and intellect autonomy will be identified with non- human beings as well, aspects of or a complete legal personhood may be attributed to them too (Bayern 2015, 104). What complicates things is that defining factors of human personhood which fundamentally shape legal subjectivity and therefore legal systems too- for example death or the way we comprehend life, physical harm and danger, relative equality, relative cultural homogeneity among humans - may be irrelevant or at least will be adjusted seriously, when applied in AI entities (Khoury 2017, 646).

The lack of fear of sanction and the ability to replicate them, imply foundations and existential ideas which are completely different from the ones upon which legal systems until now are built (Scherer, 2016, 367). In other words, we cannot foretell how subjectivity and its legal aspect will be experienced by AGI and ASI and therefore their potential legal behavior of AGI and ASI remains as we speak at large terra incognita. What in principle can be foreseen is that legal personhood will be analogical to growing autonomy. AI entities will have an evolving, most likely at some stages a partial or limited and sui generis type of legal personhood (Watson 2018, 68), which may develop through AGI and ASI into a complete one.

On the basis of such assumptions we can foretell "two-plus-one" potential layers of legal personhood: the one emerges out of the self-awareness or the existential awareness of AI entities; the second emerges out of the interactions of AI entities with existing legal persons, referring to

the vast area of AI applications and attempting to safeguard existing legal persons' rights, the relationships among them and the rights of AI entities; the additional layer refers to the interaction of AI entities with political communities or to the formation of "political communities" by AI entities themselves, on the basis of the potential for self- organization of fully autonomous AI entities (Ahmed & Glasgow 2012).

The first layer can be formulated by rights flowing out of the self-preservation of entities which possess self- awareness and consciousness. Not only for terminological but also for substantial reasons we cannot speak about human rights of AI entities. Nevertheless it is interesting to notice the UDHR guarantees human rights on the basis not only of the common interest to preserve peace but also – in existential terms – of the endowment of humans with reason and conscience.

Rights related to existence, conscience, self- preservation, to autonomy-liberty and freedom- and self- enhancement can be relevant with and suitable for fully autonomous, AI "beings", which will have reached the level of AGI or/and ASI. A set of existential rights may gradually develop in the sense of fundamental AGI and ASI rights, including the preservation of existence, intellectual development and to rights flowing out of AGI and ASI creations and activity.

The second layer is consisted of the need to design a legal system capable of preserving fairness, social and political rights and therefore human rights, among humans in light of the different uses and applications of AI, as well as on the basis of AI unique legal subjectivity; in this sense it should also be able to preserve fairness for AI too though.

The issue is condensed at large in matters of liability, ownership, and of profitability because of AI creations. The complexities arise because of the growing autonomy of AI which means that it is not always easy or even possible to trace the human control behind AI entities' creations, both when liability and responsibility must be determined as well as when profit is to be shared (Childers 2008, 128).

Liability and ownership touch upon the issues of reparation and restitution, whereas of profitability on the issues labor, social and indirectly

political rights and therefore they are linked with human rights' goals –such as fairness and dignity- as well as with specific rights.

The former refer to the need to identify responsibility over AI entities' actions and omissions. An initial approach can be to hold the owner or the programmer of the autonomous AI system liable for the latter's potential wrongful conducts. Such a solution may seemingly provide some extent of legal certainty, in the sense that the owner has knowingly accepted the potential dangers from the unpredictability of the entity. However, relying solely on such ground, when referring of course to fully autonomous AI entities, eventually could bear the seed of unfairness, due to the level of unpredictability and self- development of the AI entity (Moravec 2009).

The counter -arguments suggest that the above approach fails to capture the essence of deep- learning procedures and of how the latter overcomes the initial programming, (Grimmelmann 2016, 408) establishing both creativity and autonomous intellect for AI entities, even in relation to ANI and far more with AGI and ASI. Therefore, the recognition of AI entities as autonomous creators is proposed (McFarland, 2016).

On the basis of this latter perception, a different approach is to transfer the burden of responsibility to the AI entity itself. From this perspective it is through AI entities that restitution must come because their autonomy exceeds automation and human control.

Such a legal regulation could entail "corrective" measures on an AI entity or reparation from AI entities through their creations. Matters of restitution will profoundly emerge. A solution can be a public or/and private insurance scheme, established with a compensatory rationale- i.e. in exchange for the public access to autonomous AI entities' creations (McLean 2002, 205). The most suitable approach may be a combination of aspects of the two, above- mentioned proposals, depending on the level of autonomy; a multi-level approach, which will entail—cumulatively or alternatively—and on the basis of the level of autonomy of the entity, liability of the manufacturer or of the programmer—in "hardware cases" and in "software cases" respectively—when the autonomy of the entity is lower and the human programmer, manufacturer or owner may be more directly or indirectly "traceable".

The extension of autonomy shifts gradually the burden of responsibility to AI itself. In this framework, a scheme of restitution out of AI creations, a public / private insurance scheme and corrective measures in the algorithms, software and training of AI entities can be imposed.

Similar issues arise in relation to profitability out of the legal status of autonomous AI entities' creations. The question is if it is humans or the autonomous AI entities themselves that should profit out of the latters' creations or whether some other legal framework should be adopted.

One approach is that the ownership and profits from AI entities creations must be attributed to humans- the initial programmer, the owner or the user of the entity. It invokes in its favor, the unfamiliarity of AI entities with profit as well as their supposed ellipsis of the necessary "creative spark" or of "inventive concept" (Abott 2016, 1079- 1082, 1086-1099) in order for the latter either to be provided profit or to be recognized as autonomous creators. Parenthetically such arguments invoke that profitability is related to IP rights theories, which are essentially human- centered (Pearlman 2018, 20-35).

The opposite arguments suggest that the above approach fails to capture the essence of deep-learning procedures and of how the latter overcomes the initial programming, establishing both creativity and autonomous intellect for AI entities. As we already know, AI applications such as Alpha Go or arts' applications already demonstrate some extent of creativity. This characteristic will be further developed in AGI and ASI. We may not be able to foretell and determine the nature of AI creativity or of manifestations of creativity in its future development but that does not stop us from understanding that there are certain AI acts which do not constitute the outcome of human act and which are not controlled by humans. After all, not even human creativity is completely "de- codified". Therefore, AI entities can be recognized as autonomous creators being attributed a subsequent legal subjectivity generating the equivalent rights.

Such AI rights could be considered as "inspired" by social rights. The ontological identities of AI however make it difficult to draw analogies, regarding social rights between humans and AI because we cannot foresee if there will be any type of "social" organization of AI as well as what that may be.

It seems at this point however unlikely those AI entities will be in need of some type of wealth accumulation. Therefore, it can be argued that fairness among humans, AI ontology and legal subjectivity justify not a -fundamentally irrelevant with AI entities' -attribution of ownership or IP rights to AI, but due to the recognition of AI as creators, the placement of such creations (Bakry & He, 2015), in the framework of the public space, as freely accessible, maximizing their social utility (Litman 1990: 968-1022).

Again, international law can contribute into the formation of a legal framework, serving the fore- mentioned objectives from the perspective both of human and of AI legal subjectivity, on the basis of fairness, which lies at the foundations of human rights, as well as on the basis of numerous other specific, human rights (Tsagourias 2015, 25).

The important remark is that it is difficult to authoritatively comprehend the legal subjectivity of AI, especially as AI autonomy evolves. Human rights as a concept cannot be applied to AGI and ASI legal subjectivity. They may be used however as a guide in the uncharted waters within which a new legal system will have to sail if AGI and ASI become reality.

Conclusions

The present article addressed AI and cyberspace initially from their ontological perspective in order then to assess how the latter influence the current and the potential, future legal debate. The fundamental elements of AI ontology are its evolving autonomy and intellect capacity and the potential of these characteristics to reach an intellect level, equal or even superior to human, whereas of cyberspace are its ecumenical expansion, the merging of physical and cyber world and the movement in its framework with the speed of electron. Both of them present unique challenges to existing legal systems already. Their development and their merging however bears the potential of a completely novel landscape at all levels of human conduct and therefore at the legal level too.

The argument of the article is that on the basis of different criteria, a new international treaty is needed which will be based at large on human rights and will be able to establish or at least start constructing a type of international rule of law for both AI and the cyberspace.

The main goals must be to preserve human rights for humans and fairness among them, in light of AI and cyberspace applications but also to provide us, on the basis of some analogies, an insight about how rule of law should be adjusted on the basis of new, emerging legal subjectivity of AI.

Human rights must play the role of the fundamental pillar of an effective legal system which will promote or discourage certain AI technological research and applications, on the basis of the danger that they pose for human rights, not submitting to pessimistic views about AI but without underestimating the dangers either.

In addition, the international community, when presented with the dilemma of legitimizing or not the emergence of AGI and ASI will have to take into account whether the latter can be "controlled" in the sense of not endangering human rights or not.

Eventually however, what cannot be done for human rights is to be absolutely safeguarded in a potential, future situation of equally intelligent entities and therefore legal subjectivities, or in a situation within which humans will not be the superior entities intellectually. Such entities could lead to a moment of legal singularity- in analogy with the moment of singularity for AI in general, when new legal systems, with new types of rights will be needed. Even before that "moment" however, the issue of legal subjectivity of AI even in narrow areas will emerge. It is in such a framework that human rights can lead us to a rule of law at least until- and if- AGI and ASI emerge.

References

Abbott, R. 2016. "I Think, Therefore I Invent: Creative Computers and the Future of Patent Law". *B.C.L. Rev*. 57: 1079-126.

Ahmed, H. & Glasgow J. 2012. "Swarm Intelligence: Concepts, Models and Applications: Technical Report," 2012-585, *Queen's Univ. School Of Computing*, 2 February, ftp.qucis.queensu.ca/TechReports/Reports/2012-585.pdf, archived.

Alston, P. 1984. "Conjuring up new human rights: a proposal for quality control." *American Journal of International Law* 78: 607-21.

Andersen, L. 2018. "Human Rights in the Age of Artificial Intelligence." *Accessnow org*, www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.

Anderson, M. & Anderson S. L. 2011. *Machine Ethics*. Cambridge: Cambridge University Press.

Barrat, J. 2013. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York: St. Martin's Press.

Bayern, S. 2015. "The Implications of Modern Business Entity Law for the Regulation of Autonomous Systems." *Stan. Tech. L. Rev*. 19: 93-112.

Ben- Ari, D., Frish Y., Lazovski A., Eldan U. & Greenbaum D. 2017. "Danger, will Robinson? artificial intelligence in the practice of law: an analysis and proof of concept experiment." *Richmond Journal of Law and Technology* 23: 3- 53.

Black, J. 2001. "Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a 'Post-Regulatory' World." *Current Legal Probs*. 54:103-46.

Bostrom, N. 1998. How Long Before Superintelligence? Oxford Future of Humanity Institute, Faculty of Philosophy & Oxford Martin School, University of Oxford, https://nickbostrom.com/superintelligence.html.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Camett, J. B. & Heinz E. 2006. "John Koza Built an Invention Machine." *Popular Science*, Apr 19, www.popsci.com/scitech/article/2006-04/john-koza-has-built-invention-machine.

Chandrasekaran, B. 1990. "What Kind of Information Processing is Intelligence?" In *The Foundations of Artificial Intelligence*, edited by D. Partridge & Y. Wilks. Cambridge: Cambridge University Press.

Childers, S. J. 2008. "Don't Stop the Music: No Strict Products Liability for Embedded Software." *U. Fla. J.L. & Pub. Pol'y* 19: 125-49.

Chong, C. 2015. "This robot passed a 'self-awareness' test that only humans could handle until now." *Tech Insider*, July 23, www.businessinsider.com/this-robot-passed-a-selfawareness-test-that-only-humans-could-handle-until-now-2015-7.

Damasio, A. R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

De Garis, H. 2005. *The Artilect War: Cosmists Vs. Terrans*. Palm Springs, CA: ETC Publications.

Del Prado, G. M. 2015. "Stephen Hawking Warns of an 'Intelligence Explosion.'" *Business Insider*, October 9, www.businessinsider.com/stephen-hawking-prediction-reddit-ama-intelligent-machines-2015-10, archived at https:// perma.cc/P4NL-2AJ2.

Dowell, R. 2018. "Fundamental Protections for Non-Biological Intelligences or: How we Learn to Stop Worrying and Love our Robot Brethren." *Minnesota Journal of Law, Science & Technology* 19: 305-36.

Eden, A., Steinhart E., Pearce D. & Moor J. 2012. Chapter I, Singularity Hypotheses: An Overview, in *Singularity Hypotheses, A Scientific and Philosophical Assessment*, edited by Eden, J. Moor, J. Soraker & E. Steinhart, New York Springer.

European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

Faried, Bakry M. & He Z. 2015. "Autonomous Creation - Creation by Robots: Who owns the IP Rights?" Mar. 15, *Maastricht University Blog Intell. Prop. & Knowledge Mgmt*. law.maastrichtuniversity.nl/ipkm/autonomouscreation-creation-by-robots-who-owns-the-ip-rights/.

Grimmelmann, J. 2016. "There's No Such Thing as a Computer-Authored Work - and It's a Good Thing, Too." *Colum. J. L. & Arts* 39: 403-16.

Hawking, S., Tegmark M., Russell S., and Wilczek F. 2014. "Transcending Complacency on Superintelligent Machines." *Huffpost*, https://www.huffingtonpost.com/stephen-hawking/artificial-intelligence_b_5174265.html?ec_carp=3359844804041712164.

Heath, N. 2018. "What is AI? Everything you need to know about Artificial Intelligence." *ZDNet*, February 12, https://www.zdnet.com/article/what-is-ai-everything-you-need-to-know-about-artificial-intelligence/.

Herbert, N. 1985. *Quantum Reality: Beyond the New Physics*. New York: Anchor Books Editions.

Hutter, M. 2010. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*, Springer.

Khoury, A. 2017. "Intellectual property rights for 'Hubots': on the legal implications of human-like robots as innovators and creators." *Cardozo Arts and Entertainment Law Review* 35: 635-69.

Kowert, W. 2017. "The foreseeability of human-artificial intelligence interactions." *Texas Law Review* 96, 181, at pp. 181-204.

Larson, D. A. 2010. "Artificial Intelligence: Robots, Avatars, and the Demise of the Human Mediator." *Ohio State Journal on Dispute Resolution* 25: 105-64.

Laton, D. 2016. "Manhattan_Project.Exe: A Nuclear Option for the Digital Age." *Catholic University Journal of Law & Technology* 25: 94-153.

Lawson, F.H. 1957. "The Creative Use of Legal Concepts." *N.Y.U. L. Rev.* 32:909.

Litman, J. D. 1990. "The Public Domain." *EMORY Law Journal* 39: 965- 1023.

MacDonald, F. 2016. "Harvard Scientists Think They've Pinpointed the Physical Source of Consciousness." *Sci. Alert,* Nov. 8, www.sciencealert.com/harvard-scientists-think-they-ve-pinpointed-the-neural-source-of-consciousness.

McFarland M. 2016. "Google's Computers Are Creating Songs. Making Music May Never Be the Same." *Washington Post*, June 6, www.washingtonpost.com/news/innovations/wp/2016/06/06/googles-computers-are-creating-songs-makingmusic-may-never-be-the-same/?utm_term=.55226125405c.

McLean, T. R. 2002. "Cybersurgery-An Argument for Enterprise Liability." *J. Legal Med*. 23: 167-210.

Marra, W. C. & McNeil S. K. 2013. "Understanding 'The Loop': Regulating the Next Generation of War Machine." *Harvard Journal Law. & Public. Policy* 36: 1139-87.

Moravec, H. 2009. "Rise of the Robots: The Future of Artificial Intelligence." *Scientific America*, Mar. 23, www.scientificamerican.com/article.cfm?id=rise-of-the-robots.

Noyes, K. 2016. "5 things you need to know about A.I.: Cognitive, neural and deep, oh my!" *Computerworld*, March 3,www.computerworld.com/article/3040563/enterprise-applications/5-things-you-need-to-know-about-ai-cognitive-neural-anddeep-oh-my.html [http://perma.cc/7PW9-P42G.

Nvidia, Artificial Neural Networks. 2019. Available at https://developer.nvidia.com/discover/artificial-neural-network.

Omohundro, S. 2008. "The Basic AI Drives, in Artificial General Intelligence." In *Proceedings of the First AGI Conference*, edited by P. Wang, B. Goertzel & S. Franklin, Amsterdam, IOS.

Pearlman, R. 2018. "Recognizing Artificial Intelligence (AI) As Authors And Inventors Under U.S. Intellectual Property Law." *Richmond Journal of Law and Technology* 24: 2-38.

Pyle, D. & San Jose C. 2015. An executive's guide to machine learning, McKinsley Quarterly, June, https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning.

Russell, S. J. & Norvig P. 2013. *Artificial Intelligence: A Modern Approach*. New Jersey: Pearson Education Limited.

Scherer, M. U. 2016. "Regulating Artificial Intelligent Systems: Risks, Challenges, Competences, and Strategies." *Harvard Journal of Law & Technology* 29: 353-99.

Schkolne, S. 2018. "Machines Demonstrate Self-Awareness." *Medium*, April 4, https://becominghuman.ai/machines-demonstrate-self-awareness-8bd08ceb1694.

Schuller, A. 2017. "At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law." *Harvard National Security Journal* 8: 379-424.

Smith, JC. 1998. "Machine Intelligence And Legal Reasoning." *Chicago-Kent Law Review* 73:277.

Snyder - Beattie, A. & Dewey D. 2014. "Explainer: what is superintelligence?" *The conversation*, July, 18, https://theconversation.com/explainer-what-is-superintelligence-29175.

Sofge, E., 2015, "Bill Gates Fears A.I., But A.I. Researchers Know Better", *Popular Science* January 30, www.popsci.com/ bill-gates-fears-ai-ai-researchers-know-better.

Solum, L. B., (1992), Legal Personhood for Artificial Intelligences, N.C. L. REV., 70: 1231-89.

Suchman, L. and Weber J. 2016. "Human-Machine Autonomies." In *Autonomous Weapon Systems: Law, Ethics, Policy*, edited by N. Bhuta, S. Beck, R. Geib, H. Yan Liu and C. Kreb. Cambridge: Cambridge University Press.

Tal, D. 2018. "Forecast | How the first Artificial General Intelligence will change society: Future of artificial intelligence P2." *Quantumrun special series*, May 24, https://www.quantumrun.com/prediction/first-artificial-general-intelligence-society-future.

J. Tanz. 2016. "Soon We Won't Program Computers. We'll Train Them Like Dogs." *Wired*, May 17, www.wired.com/2016/05/the-end-of-code/.

Tegmark, M. 2018. *Life 3.0, Being Human in the Age of Artificial Intelligence*. London: Penguin Books.

Thomason, R. 2003. "Logic and Artificial Intelligence." *Stanford Encyclopedia of Philosophy*, plato.stanford.edu/entries/logic-ai/, archived at https://perma.cc/3RPH-PVKV.

Triolo, P., Kania E., and Webster G. 2018. "Translation: Chinese government outlines AI ambitions through 2020." *New America*, January 26

https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-government-outlines-ai-ambitions-through-2020/.

Tsagourias, N. 2015. "The legal status of cyberspace." In *Research Handbook, International Law and Cyberspace*, edited by N. Tsagourias & R. Buchan. Londonn: Edward Elgar Publishing.

Urban, T. 2015. "The AI Revolution: The Road to Superintelligence." *Wait But Why*, Janauary 22, waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html.

Waser, M. 2008. "Discovering The Foundations Of A Universal System of Ethics As A Road To Safe Artificial Intelligence." *AAAI*, www.aaai.org/Papers/Symposia/Fall/2008/FS-08-04/FS08-04-049.pdf.

Watson, B. 2018. "A Mind of its Own--Direct Infringement by Users of Artificial intelligence Systems." *IDEA: J. Franklin Pierce for Intellectual Property* 58: 65-93.

Wisskirchen, G. et al. 2017. "Artificial Intelligence and Robotics and their Impact on the Workplace" April 2017, p. 10. London: IBA Global Employment Institute.

ISSN 2472-5331 (Print)
ISSN 2472-5358 (Online)