IHS Working Paper 6
April 2019

Artificial Intelligence: Socio-Political Challenges of Delegating Human Decision-Making to Machines

Robert Braun

Institut für Höhere Studien
Institute for Advanced Studies
Vienna

Author(s)

Robert Braun

Editor(s)

Beate Littig

Title

Artificial Intelligence: Socio-Political Challenges of Delegating Human Decision-Making to Machines

Institut für Höhere Studien - Institute for Advanced Studies (IHS)
Josefstädter Straße 39, A-1080 Wien
T +43 1 59991-0
F +43 1 59991-555
www.ihs.ac.at
ZVR: 066207973

ARTIFICIAL INTELLIGENCE: SOCIO-POLITICAL CHALLENGES OF DELEGATING HUMAN DECISION-MAKING TO MACHINES

ROBERT BRAUN[1]

Abstract: Artificial intelligence is at the heart of current debates related to ethical, social and political issues of technological innovation. This briefing refocuses attention from the techno-ethical challenges of AI to artificial decision-making (ADM) and the questions related to delegating human decisions to ADM. It is argued that (a) from a socio-ethical point of view the delegation is more relevant than the actual ethical problems of AI systems; (b) instead of traditional responsible AI approaches focusing on accountability, responsibility and transparency (ART) we should direct our attention to trustworthiness in the delegation process; and (c) trustworthiness as a socio-communicational challenge leads to questions that may be guided by a responsible research and innovation framework of anticipation, reflexivity, inclusion, and responsiveness. This may lead to different questions policymakers and other interested publics may ask as well as novel approaches, including regulatory sandboxes and other measures to foster a more inclusive, open and democratic culture of human-ADM relations.

Key words: AI, arithmetic decision-making, delegation, Arendt, RRI.

JEL codes: M14; M31; O31; O32; O33.

Acknowledgement: *I would like to thank my colleagues Johannes Starkbaum, Tamara Brandstätter, Thomas König, Helmut Honigmayer, the participants of our regular Techno Science and Societal Transformation research group seminar and the anonymous reviewer(s) for their valuable comments to earlier drafts of this working paper.*

[1] Institute for Advanced Studies, Vienna
Techno Science and Societal Transformation Research Group

1. Introduction

Artificial Intelligence (AI) is currently seeing major media and popular interest, significant attention from regulatory and policy making bodies both on the national and on the European level, from academia and from society at large. A term coined by at Dartmouth mathematics professor John McCarthy, artificial intelligence as a research discipline was initiated at the Summer Research Project of 1956 (McCarthy, Minsky, Rochester, & Shannon, 2006). The field of AI research was launched not by agreement on methodology, choice of problems or general theory, but by the shared vision that computers can be made to perform specific tasks that may be termed as 'intelligent'. This vision was stated boldly in the proposal for the 1956 conference: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 2006, p. 12). The study and research in AI from the beginning, beyond problems of mathematics or control and operations theory, embraced the idea of duplicating human faculties such as creativity, self-improvement, and language use. AI also remained a branch of computer science that experiments with building machines that will function autonomously in complex, changing environments. From its conceptualization AI operated at the confluence of technology and society and required transdisciplinary approaches to tackle the multifold and multilayered challenges, both computational and social, that it posed.

From its origins in the 1950s, to early optimistic predictions of its founders and to recent ethical and economic challenges as well as the growing awareness of the potentially transformational nature of development, AI has seen its share of ups and downs in public interest. There has been steady growth in the past 50-60 years in basic AI research, the availability of massive amounts of data, and vast advances in computing power have brought us to a unique phase in AI history. It is now up to society in general to shape the future of the

developments and the potentially major societal impacts of artificial intelligence. While artificial intelligence became the generally accepted name of the field to signal its separation from other computational fields, it has been challenged from the outset. AI is a fuzzy term that encompasses a wide range of controversial ideas invoking connotations of human-like autonomy and intentionality as well as an 'artificial-natural' dichotomy. AI involves ideas about the possibility of machines thinking and acting humanely and/or acting and thinking rationally as well as a number of other key issues, besides those of computer science, ranging from the philosophy of the mind to probabilistic mathematics, from decision theory studied in economics to complex issues in neuroscience, psychology and linguistics (Perez, Deligianni, Ravi, & Yang, 2018).

As this working paper concentrates on the societal and human aspects of the wide field of AI it will focus on 'arithmetic decision-making' (ADM) as opposed to artificial intelligence (AI) in general. Arithmetic decision-making is part of the fuzzy field of artificial intelligence research: it deals with the challenges of data-driven machine learning-based algorithms tackling complex problems (Gillespie, 2014; Lepri, Oliver, Letouzé, Pentland, & Vinck, 2017; Willson, 2016). This limitation better describes the main social problems at hand. Arithmetically controlled decision-making is a procedure in which decisions are partly or completely *delegated* – via other persons or corporate entities – to automatically executed decision-making models to perform an action. This action, then, will have deep social implications from hiring, firing and substituting human labor to addressing social contingencies in mechanized movement on the road and, potentially, in the air, just to name a few applications being researched as we speak. Arithmetic decision making is used in setting up schemes to aid the unemployed, to design robotic vacuum cleaners, to create face recognition systems and many other technologies that are currently being experimented with.

AI is perceived to make ‘intelligent’ decisions modeled on the human mind. This poses a number of social, legal and philosophical challenges; especially as such decisions involve actions that contain ethical assumptions and moral consequences. The most well-known example of the ethical questions posed to AI is the application of the trolley problem to autonomous vehicles driven by AI. The trolley problem is the ethical model of moral decision-making when a runaway tram may kill one person or five on a railway depending on the choice of the onlooker to potentially alter its way (Bonnefon, Shariff, & Rahwan, 2016; Foot, 1978; Goodall, 2014; Thomson, 1985)). In its basic version the model describes a situation in which a tram gets loose and reaches an intersection where its route may be altered. If it continues it kills five people tied to the rail, whereas if a switch is applied the tram changes track and on that track kills one person tied to the track. There are other versions with options to stop the tram by pushing someone on the track, or deciding between tracks based on different characteristics, not the number of people to be killed by the choice. The dilemma(s) may be resolved by choosing different outcomes based on the ethical theory one subscribes to. The trolley problem points to the moral dilemma of AI: there is no one right answer, as there is no one ethical theory or moral outcome that may be applied to any specific case or circumstance. AI or arithmetic decision making in general cannot be ‘ethical’: certain decisions may be made according to specific ethical theories that work (morally) better (or are more acceptable to a community) in certain cases than do other potential choices. These technologies are also parts of socio-technical systems, such as autonomous mobility-as-a-service provision or autonomous stock trading methods, health diagnostics arrangements or corporate sales and marketing endeavors that offer complexities, both moral and social, that there is no linear or path determined way to tackle.

2. Framing arithmetic decision-making

In the plethora of politico-social challenges related to AI our attention here will focus on the *delegation of human decisions* to data-driven, arithmetically controlled systems. The delegation is part of a socio-technical framework in which ADM is utilized. This includes a delegation process, a decision-making model, an algorithm that translates the model into a computational code, the sets of data and its modes of collection the code uses as input, the 'learning', 'processing' and 'analysis' procedure the code utilizes and the entire social, economic and political environment that surrounds the delegation process. Also, the consideration and decision to apply ADM for a certain purpose, its development, procurement and deployment are parts of the framework to be analyzed.

I will look at ADM from a techno-social perspective, and discuss the social context of ADM technology as well as the techno-political consequences of delegating human decisions to machines. Questions of politics have long been part of science and technology studies (STS), and STS scholars have carefully studied how technology became intertwined with politics (Brown, 2015). Here I will focus on the socio-political challenges of delegation and discuss how research and innovation in ADM should approach these challenges as well as what are the options of regulating ADM and the process of delegation. Delegation is this context is not primarily a question of technology, but a techno-political problem that relates to the socio-technical 'imaginary' as well as the 'techno-political framework' of ADM. By 'imaginary' I mean a commonsense understanding of the delegation process: a framework that contains elements of a shared vision created, sustained and reproduced through rhetoric and power. Like other worldviews, imaginaries "become so deeply embedded in commonsense understandings that they are taken for granted and beyond question" (Harvey 2017, 24). This taken-for-grantedness animates technology innovation processes as well as policies, regulation and decisions on multiple levels, from individuals to governments.

The ‘techno-political framework’ refers to the complex interrelatedness of the principal (be it an individual or an organization) and the agent (the ADM operation of a technology) in the delegation process, as well as the socio-political context the delegation takes place and the multifold impact of this delegation and the actions that follow on society. Imaginary and framework are interrelated as the imaginary creates a commonsense understanding of taken-for-grantedness that using ADM is ‘better’ (whether ‘better’ is defined from eg. a neoliberal efficiency perspective (Means, 2015) or from a securitization angle (Yampolskiy, 2018)) than human decision-making is in specific contexts. This provides the rationale of delegation for the principal and suppresses concerns impacting society. This approach is exemplified in the EU communication on AI for Europe (date: 25.04.2018) stating that “[b]eyond making our lives easier, AI is helping us to solve some of the world's biggest challenges: from treating chronic diseases or reducing fatality rates in traffic accidents to fighting climate change or anticipating cybersecurity threats” (EC, 2018b, p. 1). Most communication on AI acknowledge that “[s]ome AI applications may raise new ethical and legal questions, related to liability or fairness of decision-making” (EC, 2018a), however, they also claim that addressing these ‘problems’ head on will make them manageable and our lives *can be made easier* by ADM.

Today the emphasis of ‘responsible AI’, a term used to tackle the socio-ethical challenges of ADM, is on the algorithmic decision-making itself and not the *relationship* between humans and ADM or the conditions when and if delegation should or should not take place. The confluence of taken-for-grantedness of delegation and the assumption that technologies are (always) socially fit for purpose create a fascination with the idea of AI induced technological progress. The constant search for a perceived betterment, whether of an individual or of society, is underscored by the belief in ADM: ‘solutionism’ provides the drive to both innovate and delegate. Innovation is framed as the unsolved not-*yet*-innovated (Dewandre,

2018) and the illusion that there is no barrier to what can be achieved by delegating our human decisions to ADM. This casts the present in terms of deficit: what is lacking, what is not proper and the future where 'solutions' (perfect or at least better decisions) lie. It also creates a structural dissatisfaction coupled with a deeply anchored impatience. This impatience, then, is the driver of delegation, but also the reason to put aside societal considerations in exchange of ADM solutionism.

While marginally related, this impatience is exemplified for instance in the current enthusiasm in transforming our education system to offer more STEM (science, technology, engineering and mathematics) as opposed to liberal arts and social science (Benson, 2017). The potential reordering of our systems of education may seriously inhibit our abilities to tackle the social challenges, ethical concerns and humane aspects of innovation in ADM and beyond. Exclusive STEM education (without social or humanities awareness) will create 'ADM subjects' -- rational principals indoctrinated by enthusiasm for technological fixes and betterment via engineering solutionism[2]. Ethical or social alarms as well as delegation concerns will be removed; ADM subjects will learn to 'think' like machines and machines will learn to think like STEM humans. One should be reminded that dystopias are not created by making machines that take over humans, but by creating humans that are not capable of controlling machines as all skills required for this control (ethics, social and political awareness) are unlearned.

Taken-for-grantedness permeates public discourse on an 'expert' and on a commonsense level as communities of power sharing similar epistemological assumptions make solutionism-visions collectively held and institutionally stabilized (Jasanoff & Kim, 2015).

[2] This is manifested in the findings of our research looking at H2020 funding principles in specific technological innovation program lines. Cf.: Akca Prill, Melek, Lindner, Ralf, Allinger, Matthias, Bernstein, Michael J., Bratan, Tanja, Braun, Robert, Gianni, Robert, Goos, Kerstin, Ikonen, Veikko, Schrammel, Maria, Nieminen, Mika, Seebacher, Lisa Marie, Tumbrägel, Tessa, Tyynelä, Janika and Wunderle, Ulrike (2018) *New HoRRIzon Deliverable 4.1. Diagnosis: RRI in Societal Challenges.* [Research Report] 256 p. http://irihs.ihs.ac.at/4918/

We are witnessing techno-optimists and the industrial-innovation complex formulating AI enthusiastic epistemic communities of power (Antoniades, 2003) acting in concert to create the imaginary of AI inevitability in all possible domains of life. This working paper entertains the idea that the moment as AI emerges from the research labs to be applied in commonly used sociotechnical artifacts (vehicles, health diagnostic instruments and commercial systems) generates the urgency to scrutinize the overall taken-for-grantedness of delegation. This paper refocuses attention from asking the question 'How can ADM (made to) be better?' to the complementary question of 'How to make our *decisions better* on when, why and how to delegate human decisions to ADM?'

The remainder of the paper will first briefly discuss how delegation is dealt with in political science, business and marketing literature. Next I will turn to current socio-political challenges to ADM, and critically discuss the responsibility, accountability, transparency framework used by 'responsible AI' scholars to address these challenges. Reframing responsibility as trustworthiness, I will show how a trustworthy ADM approach may be more appropriate to address the political problems related to ADM. It is argued that responsible AI focuses on epistemic ideals and procedures of ADM, while trustworthy ADM delegation concentrates on the socio-political context of delegation, the plurality and diversity of potential outcomes as well as the risks involved in the process of delegation. The paper will end in discussing the relevance of this to changing regulatory practices as well as draw some wider political conclusions related to our technofutures.

It is clear that the questions of ADM betterment and improving our judgment on delegation are interrelated, however not straightforwardly following from each other. The first question, important in its own right, focuses on the socio-technical appropriateness of the arithmetic decision making system as such, the second question concerns our relationship to ADM as well as the consequences of this relationship (embodied in the delegation) on others. While

the first question addresses the 'epistemic' and 'moral' qualities of ADM (and our judgments about these epistemic and moral qualities), the second focuses on the political aspects of ADM: the embeddedness of ADM in societies and our relationship to each other. The core (political) challenge is this: What our *polis* will look like and how we will deal with the socio-ethical problems when we delegate many (or most) of our socially relevant human decisions (from mobility to health and beyond) to machines with ADM.

3. Delegation

There is a wide array of research available in different areas about impacts, challenges and rationale of delegation in political science, marketing and management studies (Aggarwal & Mazumdar, 2008; Bell & Bodie, 2012; Sengul, Gimeno, & Dial, 2012). Some of the findings may be applicable to our inquiry in delegating human decisions to ADM. In political science legislative delegation to the bureaucracy is criticized as diminishing electoral accountability and exacerbating legislative shirking. In electoral theories of delegation the general consensus is that optimal political representation consists of a mixture of the delegate (no discretion) and trustee (full discretion) models of representation. One of the most important findings is that whatever model of delegation (ally, credible commitment, strategic, etc.) is applied, such models do not take into consideration the decisional as well as the social context of delegation. Decisional contexts involve the fact that principals delegate with an ongoing authority and not design their delegation strategies on a blank slate, while agents also have some already existing level of agency or de facto authority (Bendor, Glazer, & Hammond, 2001). In management literature the principal-agent problem (a version of delegation of authority in a corporate arrangement) is seen as resulting in moral hazard and conflicts of interest (related to the costs of controlling the agent and the potential benefits thereof, and/or the cost accrued by the agent to appropriately represent the interests of the principal) (Jensen

& Meckling, 1976). The principal-agent relationship generates social uncertainties tackled by regulating the relationship through individual actions, rule based processes or policy and regulatory attention. In traditional principal-agency setups agents are motivated to comply with the interests of principals through incentive mechanisms. In recent research on the behavioral aspects of the principal-agent problem principal side overconfidence as behavioral bias emerged as expected utility destructive (de la Rosa, 2011). In marketing literature delegation of consumer choice is seen as reducing felt responsibility for negative choice outcomes, and is mainly used when surrogates are perceived as knowledgeable and trustworthy. (Aggarwal & Mazumdar, 2008)

At this point it suffices to say that the problem of delegation of decisions under risk and uncertainty by differing actors is not new, therefore there is no need to reinvent the delegation-theory wheel. We may also see that human decision-making are complex setups of personal, societal and cultural determinations individualized by context, personal weighing of expected utility and a number of biases and heuristics. There is no one solution or applicable theory that says it all, therefore the analysis of delegation needs to focus on context and available delegation options as well. The core of delegation, the principal-agent problem, is not addressed via assessing the epistemic qualities of the agent, but by her intent to represent the interests of the principal suitably. This marks a shift from epistemology to a political understanding of delegation. Principal side overconfidence and the consequences thereof remind us to pay special attention to brashness *aka* taken-for-grantedness.

4. Political and ethical challenges of delegation in ADM

As advances in decisions to delegate human decisions to ADM take place at a growing pace a set of questions arise related to social, economic, political, technological, legal and philosophical issues (Dignum, 2017). Beyond ethical considerations, societies have not yet worked out ways to deal with developing new technologies that utilize machine learning,

multi-dimensional connectivity, multi-layered data collection to fit societal concerns and expectations.

(A) Accountability, responsibility and transparency

In order for ADM to pass as 'intelligent' responsibility is argued to have to be an inbuilt feature. 'Responsible AI' theorists suggest that AI systems should be augmented with the principles of accountability, responsibility and transparency (ART) as driving principles of a human centered and intelligent system (Dignum, 2017). *Accountability* refers to the need to explain and justify decisions and actions to users and other stakeholders with whom the system interacts and is ensured if decisions are derivable from, and explained by, the ADM used. *Responsibility* refers to the capability of ADM systems to answer for their decisions, to identify errors or unexpected results. This also means the need to link ADM to the fair use of data and to the actions of stakeholders involved in the system's decision. *Transparency* refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data used or created. Thus, according to responsible AI, methods are required to inspect algorithms and their results and to manage data, their provenance and their dynamics (Bostrom & Yudkowsky, 2014).

ART is problematic and may not be instructive in understanding the politics of ADM. Algorithms in ADM tend to be 'black boxes': devices that can be viewed in terms of inputs and outputs, but principals often have no knowledge of their internal workings. As algorithms that enable ADM to make decisions and navigate the complexities of their environments become more specialized and complex, even creators may no longer be able to understand them (Stilgoe, 2018). ADM is tasked to engage with complexities that cannot be captured by a set of simple and formal rules. Deep learning mechanisms operate probabilistic setups of nonlinear transformations on input to reach an acceptable level of accuracy of output. ADM

systems using such probabilities unsupervised create social uncertainties that, by design, make algorithmic decision outcomes inscrutable (Bornstein, 2016).

This is troublesome from a political ADM perspective. Politics, from a technoscience perspective, may be defined "as purposeful activities that aim for collectively binding decisions in a context of power and conflict" (Brown, 2015, p. 19). All elements of this definition sketch (human agency, purposefulness, decisions, power and conflict) require certain level of transparency related to how decisions are arrived at, what is the context of such decisions and what room the agent has to maneuver in arriving at collectively binding power setups to resolve conflict(s). Once decisions become inscrutable they may offer a technology fix to specific challenges, however the conflicts involved in finding solutions are resolved without the possibility of human agency (purposefulness) involved. This may be attractive to those who see technology as essentially 'natural' and path determined, however, from an STS perspective it annuls the well documented 'mutual constitution' of society and technology (Jasanoff, 2004). Apolitical approaches deprive technosocial situations or events of relations of power, justice, morality, and group identity as well as deny its value-laden, therefore contestable, nature. ADM thus is intertwined with a technological essentialism that abstracts technology from society.

A 'right to explanation' (Kaminski, 2018) (information about individual decisions made by algorithms) is called to help increase the accountability and transparency of the ADM process. However 'right to explanation' may be problematic as machine learning and ADM mirrors the architecture of (our current knowledge of) (Mercier & Sperber, 2017) human brains by building complex and multilayered representations of information. These representations are sometimes delusive (misreading certain information that it has mistakenly categorized) and may cause 'mental health' problems similar to hallucination or other social disorders like schizophrenia and Asperger (Hills, 2018).

(B) Trustworthiness

Trustworthiness is a key factor of interpersonal relations that enable delegation (Homburg & Stock, 2005; Sargeant & Lee, 2004). Trustworthiness may be defined as a function of assured positive anticipations in situations involving risk. Trustworthiness is generally discussed in terms of social or contractual accountability, as well as expertise involving responsibility and transparency (Keating & Thrandardottir, 2016). Some theorists of trustworthiness, however, claim that trustworthiness is not necessarily based on ART principles, but on the perceived *honesty*, *competence* and *reliability* of the other person (O'Neill, 2018). Honesty would address the claims and commitments made, competence entails the perceived ability to perform the relevant task at hand and reliability is provided to enable those communicated with to reach an intelligent judgement. Trustworthiness in this rendering is situational and communicational – it is about the institutional and social context and not the actual knowledge involved in the perceived delegated action. It is also political as it is relational: it is about assured positive anticipations of the principal about the agent in situations involving risk.

The idea of relationality is important to address the political nature of delegation. Traditional delegation frames theorize agents (both as principals and as surrogates) as 'rational' and conceptualize their relationship in terms of dominance and interest induced conflict. Delegation is based on some form of cost-benefit analysis and the implications of such analysis to the management of the relation to reach optimal outcomes. Optimal outcomes, even if social contingencies are accounted for, are framed based on starting hypotheses of short or long term interests of principals and agents, or the collective ideals of the social contexts they are embedded in. In ideal delegation setups rational agents weigh costs and benefits, analyse and adjust to contexts to arrive at utility maximizing outcomes. The relational agent, a concept originating in Hannah Arendt's conceptualization of the human condition as essentially rooted in plurality, offers an alternative route.

As opposed to the rational, the relational agent is aware of the fact that "action [is] the only activity that goes on directly between men without the intermediary of things or matter, corresponds to the human condition of plurality, to the fact that men, not Man, live on the earth and inhabit the world" (Arendt, 1958, p. 9). Plurality or relationality is a concept that challenges the modern omniscience-omnipotence utopia of 'optimal outcomes' and enables embracing diversity of the human condition both in terms of the diversity of individuals and the diversity of social contexts. It is also important that relational agents are aware of social interdependence and of the fact that there is no ultimate guarantee for a given optimal outcome as "relational selves are conscious of experiencing the 'calamities of action' arising from plurality" (Dewandre, 2018, p. 511). The calamities of action are essential to the human condition as doing away with plurality leads to the abolition of the public realm altogether (Arendt, 1958).

Politicizing ADM delegation in terms of trustworthiness means that we refocus on relational as opposed to rational agents embedded in the human condition of plurality of outcomes and the diversity of social situations. A responsible AI framed delegation seeks 'optimal outcomes' and focuses on epistemic ideals and procedures of ADM to achieve such outcomes. Trustworthy ADM delegation, in turn, focuses on the socio-political context of delegation, the plurality and diversity of potential outcomes as well as the risks involved in the process of delegation. It also embraces the diversity of potential principals as well as understands the contingencies involved in human decision making stemming from the diversity and uniqueness of identities, a basic human condition. This then allows for a more nuanced and heuristic approach to delegation, one that first looks at context, potential impacts, then rationale of delegation. Trustworthy ADM delegation also keeps the option of non-delegation open, even if delegation seems to be utility maximising on a cost-benefit basis, but utility diminishing socially or politically.

One of the socio-political contexts of machine learning in ADM to be addressed is democratizing the process of (social) learning. Machine learning advances are to be made public and shared, not to be kept proprietary to one company or technology provider. ADM applications are not only a set of engineering 'tasks' but also impact socialities that, during the delegation process, require relational awareness. This entails that addressing ADM challenges means being attentive not only to risks and challenges of the new technologies, but also to public concerns as to how and why specific innovations happen in ADM systems; why and how human decisions are transferred to machines and whether keeping human decision-making in certain contexts instead of delegating decisions to ADM would actually lead to other, alternative risk eliminating social impacts.

A good example of a trustworthy AI approach is the *European Ethics Guidelines on Trustworthy AI* created by the High Level Expert Group on Artificial Intelligence. It embraces many of the ideas including ethical purpose, stakeholder involvement and constant reflection expressed in this working paper (EC, 2018c). The guidelines focus on human-centric AI stating that should be developed, deployed and used with an ethical purpose, grounded in, and reflective of, fundamental rights, societal values and the ethical principles of beneficence (do good), non-maleficence (do no harm), autonomy of humans, justice, and explicability. It also focuses on paying appropriate attention to situations involving vulnerable groups and to situations with asymmetries of power or information. These are essential to the delegation of decision making from humans to ADM. It calls attention to incorporate data governance, design for all principles, human oversight when required, non-discrimination and respect for privacy from the earliest design phase. The document also suggests providing, in a clear and proactive manner, information to stakeholders about the system's capabilities and limitations, allowing them to set realistic expectations of the

capabilities and potentials of the ADM operated AI. This is crucially important to assess when and for what end to delegate.

However, the document falls short of mentioning that non-delegation is also an option for specific contexts and by specific stakeholders. The document calls attention to make trustworthy AI part of the organization's culture, and provide information to stakeholders on how trustworthy AI is implemented into the design and use of AI systems. The document also emphasizes to ensure participation and inclusion of stakeholders in the design and development of the AI system. The document provides an excellent overview of how, for what purpose and by which means trustworthiness may become part of AI/ADM. It comes short however, of drawing the attention to potential barriers and pitfalls of the delegation process as well as the contexts where delegation is beneficial and where it may or should be avoided.

(C) Responsible & trustworthy ADM

When applying basic principles of trustworthiness to ADM, *honesty as claims and commitments of ADM* may mean that ADM development avoids an overall technology-fix approach that looks at social challenges as requiring a better (more efficient, faster, safer) technology. These concepts are socially loaded and may lead to unintended social consequences. In case of ADM public deliberation on the specific technologies, their benefits and pitfalls, wellbeing impacts as well as their alternatives should be evaluated. Disruptive technologies, ADM included, claim to offer remedies to past social pathologies of technological development, such as inequality, social exclusion or ethical dilemmas. This may, once again, lead to solutionism.

The concept of *competence as the perceived ability to perform relevant task* means that delegation of human decisions to ADM requires new regimes of governance that do not take technology as fixed and seen to provide solutions to deficits of human behaviour,

infrastructure or the law. Technological systems should be shaped actively, including the (sociotechnical) imaginaries that animate them and (re)connecting ADM with their social and natural environment in multiple ways and forms. This would also mean that keeping non-delegation options open is also part of the assessment of competence. Involving different stakeholders early on in the innovation process would create a more open discussion about what domains and areas should ADM technology be applied to. This is extremely important as the delegation of human decision making to ADM becomes unavoidable once a technology has been developed and emerged as market ready. When considering this question, responsible innovators as well as other stakeholders – policy makers, funders and regulators in ADM should first reflect on the research and innovation process itself and openly discuss what roles ADM run artefacts should or should not play in society.

A more open, anticipatory and democratic approach to ADM would ask what different publics would want from ADM – where do they think delegating decisions may be useful. This would require that these publics have a better understanding what ADM can and cannot do, while openly conversing about the potential intended and unintended social impacts from job loss to creating lethal autonomous robots for terrorist operations. Such programs would also empower and engage the public in debates about technology futures, require substantial learning on their side to have an educated conversation. Public engagement would offer inclusion of different points of views as well as different aspects and arguments for the social desirability of specific ADM systems and their governance, including collaborative practices of co-design and co-regulation. *Reliability as enabling those communicated with to reach an intelligent judgement* would empower responsible innovation in ADM to regard humans as resource and not as problem.

Claims of ADM that they provide a solution to human errors in driving is living proof. While, for example, safety is an utmost concern of mobility, in the course of autonomous vehicle

ADM innovation, the question ‘safe enough for what?’ should also be posed. Automobility for instance traditionally ‘accepts’ roadside death and injury as part of the system implicitly applying a consequentialist ethical position in which the benefits of the car (both in mobility and in social terms) are assumed to outweigh the malign impacts of accidents, while other mobility technologies, e.g. elevators, became widespread and societally transformational only after the technology was considered to be absolutely safe and accidents were eliminated (Bernard, 2014).

Institutions and individuals need to build and develop appropriate reflexive capacity to diverge from a technology fix approach and focus on social learning, complex assessments of impacts and responsiveness to challenges thereof, both in the sense that people learn and assess impacts socially and that societies learn, reflect and respond constantly, offering ways to better understand and democratize the social experiment of ADM transition.

5. Regulatory sandboxes

The framework and practice of responsible research and innovation (RRI), a concept aiming at the betterment of the social embeddedness of research and innovation (R&I), may be applied to ADM. Responsible innovation addresses the growing distrust in scientific knowledge by challenging the established neoliberal model of R&I in which research and innovation have been understood as a linear, efficiency and progress driven development, directly leading from basic research to application, from experts to consumers. It points towards a research and innovation ecosystem that recognizes non-linearity and addresses inclusion of different publics with regards to research goals and questions, processes, outcomes and impacts. RRI, according to one definition, is "taking care of the future through collective stewardship of science and innovation in the present," which may be achieved through applying a framework of anticipation, reflexivity, inclusion, and responsiveness (Stilgoe, Owen, & Macnaghten, 2013).

Applying an anticipatory and reflective ADM assessment framework would focus on diversity of potential outcomes and the implications thereof, as well as on the potential of cooperation as opposed to delegation. Currently ADM is perceived to substitute human decision making in many areas; an alternative social vision could look at options of complementing the strength of human intelligence with ADM and other forms of AI. This would involve complex strategies from education to participatory decision-making, focusing, as discussed earlier, on options like not introducing more STEM into education but on how to utilize STEM for arts, humanities and vice versa. The confluence of different forms of arts and STEM[3] as well as the experiments of the Socio-Technical Integration Research embedding social scientists and humanities scholars in laboratories (Fischer, 2007) provided

[3] One example is the experiment, in a different context, in the Joint Research Center: cf. https://ec.europa.eu/jrc/en/event/exhibition/resonances-science-arts-politics.

experimental avenues to bring a more value and human centered design to ADM and AI in general.

From a regulatory point of view it is important to note that ADM is a technology that is being tested and deployed, as well as work-in-progress. Therefore ADM poses challenges not only to researchers but also to policy makers and regulators. One of the main challenges is how to apply an iterative, anticipatory regulatory practice to the emerging field of ADM. As opposed to advisory or adaptive regulatory practices, methods usually applied to regulation in emerging technological fields a more open, anticipatory approach may be applied (Armstrong & Rae, 2017). This would assure a better understanding of technology's impact on economy and society, as well as foresee the constantly changing regulatory needs and vision for the future. Such regulation would involve a wide array of stakeholders in the regulatory process including regulators, businesses, cross-industry, civil society, local authorities, cities, citizens, early adopters and NGOs.

In order to avoid regulatory traps of both policy push, and regulatory blockage regulatory sandboxes could be created where innovators together with citizens and other stakeholders may experiment with ADM/AI technologies, involving and engaging knowledge of diverse stakeholders so innovation in ADM may also be attentive to complex social impacts and uncertainties. The focus of regulation, besides ethical AI, should also concentrate on the delegation process and the pre-conditions thereof. In such regulatory sandboxes specifics of delegation – when, how, why delegation should or should not happen – could be experimented with, also unearthing intended and unintended social and ethical consequences. Regulators may also learn and adjust regulatory regimes as work-in-progress. This is especially important as deployment of ADM requires constant regulatory adaptation. The challenge is that adaptation may also happen through de-facto regulation, industry initiatives

being standardized through industry membership organizations or by early mover practices, not taking multiple stakeholder perspectives into account.

6. Conclusion

Ethical issues and societal challenges burgeon AI research. Within the social aspects of AI this paper focused on discussing, based on recent literature, as to when, why and how delegate human decisions to ADM. This marks a move from the epistemic and ethical challenges in AI/ADM innovation towards a societal and pluralistic understanding of the issues of ADM. ADM will become more complex and multifold, thus in most cases it will constantly improve maneuvering the complexities of decision making and decide making actions that are perceived to be safer, faster and more reliable. This is good. However, societies are complex and fuzzy and thus some of the seemingly safer, faster and more reliable decisions may not to turn out to be beneficial for society as a whole. This is the by-now-old story of unintended consequences (Blok & Lemmens, 2015).

The emergence of AI/ADM from test labs to general use artefacts means that in more and more instances we will delegate our human decisions to machines. Beyond unintended consequences this may change our social world profoundly. Autonomous vehicles, robotic diagnostic and care systems, arithmetic social platforms will dominate our worlds to make decisions and act on our behalf. The urge for delegation is animated by the taken-for-granted imaginary called progress (Benjamin, 2016). Responsible AI focuses on the epistemic and ethical ideals that may be built into the machines with ADM. Politicizing ADM starts from the relationality of the human condition and refocuses attention from the inside of the machines to the outside where they function. Outside is the social context in which ADM machines operate as well as the space between humans and machines. This is the space where delegation happens. If the *polis*, the public realm of the political where humans nourish

freedom, is space in-between people (Arendt & Schell, 2006) then we are redirecting our attention to the new technopolis of in-between of people and machines. Thus the real challenge of ADM is not ethical or epistemic but, as has also been argued in this paper, political. Every decision to substitute human decision-making with ADM shrinks the in-between of the public realm. Responsible AI would like to save the in-between and thus recreate the public realm by making machines more humane. Focusing on trustworthiness aims to tackle the political challenge head on. It tries to keep non-delegation options open by forcing us to think when delegation of decisions by humans to machines does actually make the in-between of our technopolis populated by people and machines one notch smaller.

References

Aggarwal, P., & Mazumdar, T. (2008). Decision delegation: A conceptualization and empirical investigation. *Psychology & Marketing, 25*(1), 71-93. doi:10.1002/mar.20201

Antoniades, A. (2003). Epistemic Communities, Epistemes and the Construction of (World) Politics. *Global Society, 17*(1), 21-38. doi:10.1080/0953732032000053980

Arendt, H. (1958). *The Human Condition*. Chicago: Chicago University Press.

Arendt, H., & Schell, J. (2006). *On Revolution*: Penguin Publishing Group.

Armstrong, H., & Rae, J. (2017). *A working model for anticipatory regulation*. London.

Bell, R. L., & Bodie, N. D. (2012). Delegation, Authority and Responsibility: Removing the Rhetorical Obstructions in the Way of an Old Paradigm. *Journal of Leadership, Accountability and Ethics, 9*(2).

Bendor, J., Glazer, A., & Hammond, T. (2001). Theories of Delegation. *Annual reviews of Political Science, 4*, 235-269.

Benjamin, W. (2016). *On the Concept of History*: Createspace Independent Publishing Platform.

Benson, J. D. (2017). STEM and the Humanities. *Open Access Library Journal, 4*(e3476.).

Bernard, A. (2014). *Lifted: A Cultural History of the Elevator*. New York: NYU Press.

Blok, V., & Lemmens, P. (2015). The Emerging Concept of Responsible Innovation. Three Reasons Why It Is Questionable and Calls for a Radical Transformation of the Concept of Innovation. In B.-J. Koops, I. Oosterlaken, H. Romijn, T. Swierstra, & J. van den Hoven (Eds.), *Responsible Innovation 2: Concepts, Approaches, and Applications* (pp. 19-35). Cham: Springer International Publishing.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573-1576.

Bornstein, A. M. (2016). Is Artificial Intelligence Permanently Inscrutable? *Nautilus*(40).

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press.

Brown, M. B. (2015). Politicizing science: Conceptions of politics in science and technology studies. *Social Studies of Science, 45*(1), 3-30. doi:10.1177/0306312714556694

de la Rosa, L. E. (2011). Overconfidence and moral hazard. *Games and Economic Behavior, 73*(2), 429-451. doi:https://doi.org/10.1016/j.geb.2011.04.001

Dewandre, N. (2018). Political Agents as relational Selves: rethinking EU Politics and Policy-Making with Hannah Arendt. *Philosophy Today, 62*(2), 493-519.

Dignum, V. (2017). Responsible Artificial Intelligence: Designing for Human Values. *ITU Journal: ICT Discoveries*(Special Issue No. 1,).

EC. (2018a). *Artificial Intelligence*. Brussels: European Commission Retrieved from https://ec.europa.eu/digital-single-market/en/artificial-intelligence#" name=.

EC. (2018b). *Artificial Intelligence for Europe*. Brussels: European Commission.

EC. (2018c). *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. Brussels: European Commission.

Fischer, G. (2007). *Designing socio-technical environments in support of meta-design and social creativity*. Paper presented at the Proceedings of the 8th iternational conference on Computer supported collaborative learning, New Brunswick, New Jersey, USA.

Foot, P. (1978). *The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices* Oxford: Basil Blackwell.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–193). Cambridge, MA: MIT Press.

Goodall, N. J. (2014). Machine Ethics and Automated Vehicles *Road Vehicle Automation* (pp. 93-102). London: Springer.

Hills, T. T. (2018). The Dark Side of Information Proliferation. *Perspectives on Psychological Science*, 1745691618803647. doi:10.1177/1745691618803647

Homburg, C., & Stock, R. M. (2005). Exploring the conditions under which salesperson work satisfaction can lead to customer satisfaction. *Psychology & Marketing, 22*(5), 393-420. doi:10.1002/mar.20065

Jasanoff, S. (2004). *States of knowledge: the co-production of science and the social order*: Routledge.

Jasanoff, S., & Kim, S.-H. (2015). *Dreamscapes of Modernity*. Chicago: Chicago University Press.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics, 3*(4).

Kaminski, M. (2018). The Right to Explanation, Explained. *https://doi.org/10.31228/osf.io/rgeus*

Keating, V. C., & Thrandardottir, E. (2016). NGOs, trust, and the accountability agenda. *The British Journal of Politics and International Relations, 19*(1), 134-151. doi:10.1177/1369148116682655

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology, 31*.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine, 27*(4).

Means, A. (2015). On accelerationism- decolonizing technoscience through critical pedagogy. *Journal for Activism in Science & Technology Education, 6*(1), 21.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

O'Neill, O. (2018). Linking Trust to Trustworthiness. *International Journal of Philosophical Studies, 26*(2), 293-300. doi:10.1080/09672559.2018.1454637

Perez, J. A., Deligianni, F., Ravi, D., & Yang, G.-Z. (2018). *Artificial Intelligence and Robotics*. Retrieved from London:

Sargeant, A., & Lee, S. (2004). Trust and relationship commitment in the United Kingdom voluntary sector: Determinants of donor behavior. *Psychology & Marketing, 21*(8), 613-635. doi:10.1002/mar.20021

Sengul, M., Gimeno, J., & Dial, J. (2012). Strategic Delegation: A Review, Theoretical Integration, and Research Agenda. *Journal of Management, 38*(1), 375-414.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science, 48*(1), 25-56.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy, 42*(9), 1568-1580.

Thomson, J., J. (1985). The Trolley Problem. *Yale Law Journal*, 1395–1415.

Willson, M. (2016). Algorithms (and the) everyday. *Information, Communication & Society*.

Yampolskiy, R. V. (2018). *Artificial Intelligence Safety and Security*. London: CRC Press.