Artificial intelligence: From ethics to policy

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 641.507 – June 2020

EN

Artificial intelligence: From ethics to policy

There is little doubt that artificial intelligence (AI) and machine learning (ML) will revolutionise public services. However, the power for positive change that AI provides simultaneously holds the potential for negative impacts on society.

AI ethics work to uncover the variety of ethical issues resulting from the design, development, and deployment of AI. The question at the centre of all current work in AI ethics is: **How can we move from AI ethics to specific policy and legislation for governing AI**?

Based on a framing of 'AI as a social experiment', this study arrives at policy options for public administrations and governmental organisations who are looking to deploy AI/ML solutions, as well as the private companies who are creating AI/ML solutions for use in the public arena. The reasons for targeting this application sector concern: the need for a high standard of transparency, respect for democratic values, and legitimacy. The policy options presented here chart a path towards accountability; procedures and decisions of an ethical nature are systematically logged prior to the deployment of an AI system. This logging is the first step in allowing ethics to play a crucial role in the implementation of AI for the public good.

AUTHORS

This study has been written by Dr Aimee van Wynsberghe of Delft University of Technology and co-director of the Foundation for Responsible Robotics at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

Acknowledgments

The authors would like to thank the following researchers who helped in giving constructive feedback on the ideas and content found in this study: Scott Robbins, Madelaine Ley, Tom Coggins and Acacia Parks.

ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in June 2020.

PE 641.507
ISBN: 978-92-846-5855-8
doi: 10.2861/247
QA-03-19-800-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (Intranet)
http://www.europarl.europa.eu/thinktank (Internet)
http://epthinktank.eu (blog)

Executive summary

There is little doubt that artificial intelligence (AI) and machine learning (ML) algorithms will revolutionise healthcare, logistics, human resources, policing, education, and other public services. Algorithms are already used in numerous social service contexts, including pretrial decisions for detecting risks of defendants re-offending, detection of child maltreatment, and predictive policing, which allocates police throughout the city for crime prevention. Regardless of the application domain, the power for positive change that AI provides simultaneously holds the potential for negative impacts on society.

The way in which AI/ML progress on a global, national, or international scale is dependent upon the vision put in place by academics, policy-makers, industry leaders, public administration organisations, consumer rights organisations and the like. This study is meant to show a vision of a future world that conceptualises AI as a real-world experiment and thus requires that it meet the conditions of an experiment, i.e. that it only be conducted when: 1) there are appropriate ethical constraints in place to protect citizens, 2) the experiment is aimed at assessing a predicted amount of good to be achieved by the AI/ML system, and 3) any (acceptable) risks are appropriately balanced against the assured benefits for users/society.

It is important to note that this study is not aimed at suggesting that ethics will solve the breadth of issues that arise from the design, development and/or use of AI. Instead, this study takes the concept that has been explicitly referenced in countless guidelines and corporate strategies, 'ethics', and explains **what** ethics is, and **how** ethics ought to be understood as a resource in the AI debate beyond its current use, namely to generate principles.

From an overview of the present-day ethics and technology literature, the following key insights were highlighted:

- Insight #1: Transparency of AI algorithms can mean three distinct things: first, the complexity of modern AI systems – leaving lay users in the dark; second, the intentional obfuscation by those designing AI solutions – leaving lay users and policy-makers in the dark; and third, the inexplicability regarding how a particular input or inputs result in a particular output or outputs – leaving everyone in the dark;
- Insight #2: Bias and fairness of AI/ML algorithms resulting from the training data is a significant barrier to the ethical development and use of AI/ML. Important questions concerning what is 'fair', what is 'accurate' and how to balance trade-offs between the two must be analysed; Insight #3: AI should be understood as a socio-technical system and should be assessed according to the society in which it has been created, while society's role in the development and applications of AI/ML should not be under-estimated;
- Insight #4: AI can be designed and implemented in ways that obfuscate attributions of responsibility and accountability, but that does not necessarily mean that responsibility and accountability are not possible in the context of AI;
- Insight #5: Risk assessments, while valuable, do not capture important ethical risks which may be unquantifiable, qualitative, or unobservable;
- Insight #6: Ethical technology assessments (eTA) are a viable mechanism for uncovering novel ethical issues which may arise due to the development and use of AI; and,
- Insight #7: The proliferation of AI in society is an ongoing social experiment full of risks and hypothesised benefits.

The field of AI ethics has been tasked with uncovering the variety of ethical issues resulting from the design, development, and deployment of AI. The outcome of studying these issues, for many institutions and organisations, has been to create AI ethics guidelines to inspire and steer the responsible development and use of this technology. Of late, scholars and practitioners are exploring the move 'from principles to practice' to inform the application of AI ethics principles and

guidelines in industry. The question at the centre of all the ethical issues raised, all the AI ethics principles suggested, and all the technical solutions proposed, is: **How can we move from AI ethics to specific policy and legislation for governing AI**?

This study reiterates that the issues experienced within society upon the introduction of AI/ML are critically seen through the lens of ethical reflection and concludes with possible solutions in the form of policy options for legislation. The specific policy options are directed towards the public administration and governmental organisations who are looking to deploy AI/ML solutions, as well as the private companies who are creating AI/ML solutions for use in the public arena. The policy options centre around the practice of logging and its relationship to ensuring accountability. The logging discussed here has to do with the ethical considerations relevant to the technology, and ultimately, the creation of a systematic procedure for ethical evaluations that is thoroughly documented. The ethical constraints follow directly from the key insights attributed to the ethics of technology as a field of study.

The reasons for targeting this application sector concern both the desire to use AI/ML in these spaces, along with the need for this sector to maintain a high standard of transparency, respect for democratic values, and legitimacy. The reasons to legislate are multiple: the criticality of ethical and human rights issues raised by AI/ML development and deployment; the need to protect people (i.e. the principle of proportionality); the interest of the state (given that AI/ML will be used in state-governed areas such as prisons, taxes, education, child welfare); the need for creating a level playing field (e.g. self-regulation is not enough); and the need for the development of a common set of rules for all government and public administration stakeholders to uphold.

Based on a framing of 'AI as a social experiment,' this study arrives at the following policy options, understood as ethical constraints, for European Parliamentary policy-makers:

1 It is proposed that all AI/ML system developers are required to hold a data hygiene certificate (DHC) to be eligible to sell their solutions to government institutions and public administration bodies. It is well known that the quality (or hygiene) of the data plays a key role in the efficacy and accuracy of the algorithm. Without accurate algorithms, the autonomously developed rules (of ML) will also be skewed. Consequently, a first ethical constraint is to ensure the quality of the data being used to train the algorithm, where quality is measured according to its sourcing, acquisition, diversity, and labelling. Such a certificate does not require insight into the proprietary aspects of the AI system (i.e. companies do not have to divulge their algorithm) and, of equal importance, such a certificate does not require organisations to share their data sets (which may be their source of income) with competing organisations.

2 It is proposed that all public and government organisations using AI systems are required to conduct an ethical technology assessment (eTA) prior to deployment of the AI system. The eTA is a written document intended to capture and log the dialogue that occurred between ethicist and technologist and/or ethicist and public administration officials about to implement the AI/ML solution. The eTA is a list of ethical issues related to the AI/ML application, made by an expert trained to engage in ethical reflection (or at the very least one who is able to envision possible moral risks related to the implementation of the AI/ML). The eTA is the moment where one must consider all the possible ethical risks that could result from the AI/ML application in question.

3 It is proposed that all public administration institutions and government bodies are required to show clear goals for the AI/ML application. With this policy option, AI/ML would not be deployed in society in the hope of learning an unknown 'something'. Instead, it is proposed that there must be a specific and explicit 'something' to be

learned. It is suggested that the specific aim and scope of the AI/ML experiment must also be stated as part of the eTA.

4 It is proposed that all organisations deploying AI systems should produce an 'accountability report' in response to the eTA. The accountability report is the third step in logging the AI/ML use in public administration and/or in government. Whereas the eTA is meant to draw out the possible negative consequences of implementing an AI system (completed by an external third party), the accountability report is a response to the eTA, completed by the organisation implementing the AI/ML system. It is meant as a response to the ethical and human rights issues that were identified in the eTA. Thus, in the accountability report, it is proposed that institutions will be required to account for how they have mitigated or corrected the concerns raised in the eTA.

Table of contents

Table of figures

1. Introduction

Artificial intelligence (AI) holds great power in solving some of the world's most dangerous and complicated problems. For example, AI can enable incredible energy savings in large organisations, predict forest fires and other natural or health disasters to direct resources in a timely manner, and identify patterns in healthcare data to assist in the diagnosis of patients. There is little doubt that AI will revolutionise healthcare, logistics, human resources, policing, education, and other public services. Algorithms are already used in numerous social service contexts, including pretrial decisions for detecting risk of defendants to re-offend (1), detection of child maltreatment (2), and predictive policing, which allocates police throughout the city for crime prevention (3). While its applications are vast, AI brings novel ethical challenges that threaten both users and non-users of the technology. Across the globe, we hear stories of AI-driven mishaps. In one case, AI tools mistakenly identified innocent people as criminals. In another, AI systems intended to remove bias from hiring were trained in such a way that they inadvertently developed the same biases that hiring managers showed – they gave preference to male candidates for high-level positions. Regardless of the application domain, the power for positive change that AI brings simultaneously holds the possibility for negative impacts on society. The question arises: **What can be done to minimise harm while maximising the benefits of AI solutions?**

The field of AI ethics has been tasked with uncovering the variety of ethical issues resulting from the design, development, and deployment of AI. Some work on AI ethics has focused on developer practices that create problematic situations, such as insecure data storage, use of model training practices that allow for bias to develop, and failure to disclose details of algorithm contents (see reference list numbers 1–3, 6, 7). Other work points to potentially dangerous societal impacts of AI, including human rights infringement, potential loss of control of AI growth, and differential impact within society (see references 4, 5, 7–15).

The result of studying these issues, for many institutions and organisations, has been to create AI ethics guidelines to inspire and steer the responsible development and use of this technology. To be sure, there is great strength in the creation of principles for guiding technology's development and use. In fact, the idea of applying strong ethical principles is not new. For decades, the biomedical ethics principles of autonomy, beneficence, non-maleficence, and justice have shaped clinical practice (see reference 16). Consider, for example, when a medical treatment is available to someone but their religious affiliation prohibits him/her from receiving the treatment. The principles are meant as a framework for guiding clinical decisions in such cases; physicians must balance their duty to do good with the autonomy of the patient to decide for or against treatment. Following the maltreatment of humans as subjects in medical experimentation (e.g. the Tuskegee study in the United States of America), these principles exist to prevent the mistreatment of individuals and groups in medicine; it follows, then, that a set of guiding principles might serve the AI field in the prevention and mistreatment of citizens and similarly inform its growth.

It is important to note, however, that the bioethics principles are meant as a tool to reflect on and evaluate a new technology – a new drug, or a new procedure. They are not a pre-packaged answer to the question of whether the 'new thing' is good or bad. They are not the solution to all the problems facing the patient or the healthcare practitioner. They are simply a framework for asking questions; specific mechanisms are needed to put them into action. The bioethics principles are overseen by a governing body (e.g. the US Food and Drug Administration or the European Medicines Agency) who will govern the process by which new drugs or technologies are evaluated and tested. The principles are also upheld by ethical review boards in hospitals, who deliberate what is 'best' for patients. In summary, in addition to establishing ethical principles, concrete processes and regulatory bodies are needed to realise them.

Of late, scholars and practitioners are exploring the move 'from principles to practice' to inform the application of AI ethics principles and guidelines. Some of this work is about translating ethical principles into technical requirements (see references 17–19) and/or design methodologies such as privacy-by-design to ensure privacy as a default of the system, ethics-by-design to ensure that values are deliberately included into the design process (see 20, 21), or ethically aligned design to ensure that all decision choices be aligned with ethical values (see 19, 20, 22, 23). However, the uniqueness of AI, in particular a machine learning (ML) algorithms, centres on its complexity and opacity; the rules governing an ML algorithm may be unknown to the human developer, making it difficult to align said rules with ethical values (to be discussed in detail in section 5.3). Because of this complexity and opacity, in combination with the ubiquity of AI/ML systems, the question at the axis of all the ethical issues raised, all the AI ethics principles suggested, and all the technical solutions proposed, is: **how can we move from AI ethics to specific policy and legislation for governing AI?**[1]

The governance of AI is neither about stifling innovation nor about neglecting AI's possible benefits to society; rather, it is about recognising that AI, as it is practiced now, must be understood as a real-world experiment, and this experiment needs to be regulated to protect the subjects who are involved and/or impacted. The purpose of this report is to take a closer look at the domain of ethics – ethics and technology in particular – and to ask what kind of insights and policy options can be drawn, and what policy options should be derived from these insights (about the ethics of AI) for policy-makers. In doing so, light is shed on the kinds of policy options for AI governance that can accompany a set of principles and/or guidelines.

Before going forward, however, it is important to understand that ethics is not about checking boxes. Rather, ethics is a form of deliberation, critique, and inquiry. There are a series of questions we can take from the domain of ethics as a starting point for stimulating the kinds of questions we should be asking: 'does this overly benefit certain groups over others'; 'are there unintended consequences resulting from this action'; 'what is the best thing to do in this situation'? These are examples of questions about the ethics of designing and implementing AI that do no't have an easy answer.

Second, the ethics of AI is not a one-time event; rather, it is a process of continual reflection about the technology's impact on society. New ethical concerns will arise as we learn more about how AI is built and used. Furthermore, values are not static and will change (for better and for worse) (see 24); any value-driven set of guidelines will need periodic revision to accommodate growth. Consider, for example, that only in recent months are experts beginning to study the environmental impact (e.g. the carbon footprint) and the human rights implications (e.g. slave labour-like conditions in labelling factories) of training ML algorithms. Ethically inspired regulation of AI should be able to capture new developments to ensure that the debate on AI ethics facilitates a constant reflection on new and emerging ethical concerns.

Third, the ethics of technology teaches us that AI in society, like any other new technology, should be conceived of as a social experiment. Framing the development and deployment of AI as a social experiment means having a discussion about the ethical constraints of this experiment (i.e. subjects should be asked to give 'consent' when participating in the experiment) and the kinds of things we need to learn during the experiment (e.g. how these benefits and harms are being measured and

[1] It should be noted that at the same time that this study was completed, the European Commission High-Level Expert Group on AI released policy options for regulation as an addendum to their Guidelines for Trustworthy AI. That overall exercise was not about explaining how ethics can help us arrive at regulation; rather, it was directed at how principles can be directed towards regulation. The study here is about a deep dive into what ethics is and how ethics can be translated into policy options.

studied). In other words, rather than thinking of the beneficial results of AI or the risks of AI as 'nice to know,' we must consider them as factors that society 'needs to know'.

This study is also meant to address the attack on AI ethics/ethicists as 'ethics washing'. There have been criticisms of late in the public and private sectors about ethics washing, i.e. having ethics boards with no clear mandate, or having industry determine what governments should do. But to be clear, **that is not how ethics has been practiced** to date in a variety of contexts (e.g. in healthcare ethics review boards or university ethics boards) and **that is not how ethics needs to be practiced**. There are examples to date showing how the creation of ethics boards with a clear role, mandate, task to evaluate, choice to make, or role to play succeed in creating space for the ongoing reflection and evaluation of technologies and their impact on society (or on specific demographics within society, e.g. patients). It is therefore time to decide what role ethics/ethicists can, and cannot, play in AI governance.

This study progresses as follows. We begin with a brief overview of AI as a technology and the unique features it brings to the discussion of ethics: What is AI and what is new about it that is deserving of ethical attention. Following this, we outline what the policy-maker should understand about ethics: what is ethics as a discipline and what lessons can be learned from ethics and applied to AI. Most important is translating these ethical considerations into concrete policy options for policy-makers. This study will provide greater clarity on the practical lessons that can be learned from ethics and technology, as well as the kinds of advice that ethics can provide to policy-makers. The goal of the final section is to articulate the insights with which ethics confronts policy-makers. The policy options made are meant to create basic minimal requirements for all public institutions and government organisations using AI/ML solutions to meet. these basic requirements Furthermore, this study proposes that these basic minimal requirements should be enshrined in regulations to govern the development and use of AI in public institutions across Europe. This study is not an effort to raise and solve every ethical issue;[2] rather, the task of this study is to provide a vision of how AI/ML can be governed ethically and the first steps to get there.

This document is directed at policy-makers who will provide policy options in the development, procurement, and deployment of AI to governmental and public organisations, as well as private companies creating AI products and services to be used in the public sector (as a public service). This is done because **such organisations must be subject to a higher standard of transparency in the service of democracy and the protection of human rights, the foundational elements of a liberal democracy**. By directing these organisations to make public ethical technology assessments and accountability reports prior to the deployment of AI/ML algorithms, the goal is to invite public comment and feedback. In this way, companies and public institutions alike can be pressed to consider broader public input and earn public acceptance as a criterion for legitimacy.[3]

[2] For studies and reports of that kind please refer to the following references at the end of this study: 8,14,19,25

[3] The choice to direct these policy options towards public institutions comes from their need for greater transparency and legitimacy but in no way means that private organisations should not also follow these regulations. This ought to be the next stage of policy options.

2. Objectives

The objectives of this study are to develop:

- ethically informed policy options for any AI/ML product being used in government and/or the public sphere;
- an AI ethical technology assessment framework that takes as its inputs the context, the application, and the specific AI algorithm to facilitate a targeted analysis of AI in a variety of application contexts;
- stakeholder specific policy options for the responsible implementation of AL/ML products, aligning them to defined values and ethical principles that prioritise human well-being in a given context.

3. Artificial intelligence

This section is meant to introduce the reader to the technology in question by outlining the key features of AI that make it unique. The aim is to show that AI ethics is also unique and, as such, policy options for AI governance will have distinctive features.[4]

To begin, artificial intelligence (AI) was described in its early days as 'machines that mimic 'cognitive' functions that humans associate with other human minds, such as 'learning' and 'problem solving'' (26), to imitate humans, if you will (27). Others suggest that AI brings a core component of intelligence, i.e. prediction (28). For example, 'deep Genomics improves the practice of medicine by predicting what will happen in a cell when DNA is altered. Chisel improves the practice of law by predicting which parts of a document to redact. Validere improves the efficiency of oil custody transfer by predicting the water content of incoming crude' (28)

For this study, we use the definition put forward by the European Commission High-Level Expert Group on AI:

> *Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).*[5]

According to this definition, an AI system is defined by showing a certain level of intelligent behaviour, in so far as it can act without real time human input (in other words, it can act autonomously). Moreover, AI is understood as software that can be both embodied in the world (in a robot, for example) or can exist as software-based (software in recruiting new employees, for example). This is important to remember – AI and robots are not synonymous, but are distinguished by the element of embodiment.

AI has now become an umbrella term used to discuss the variety of technologies that can mimic, automate, or outperform the capabilities of human intelligence. Some such methods are: symbolic or good old fashioned AI (GOFAI), machine learning (ML), or statistical methods. Within ML, there are further methodologies such as deep learning, neural networks, convolutional neural networks, evolutionary algorithms, etc.

ML systems are generally what the media and others are referring to when they talk about the success of AI in the last decade. Image classification (e.g. labelling skin moles as cancerous), facial recognition, and game playing (e.g. DeepMind's AlphaGO, who beat the world champion in 2017) all rely on ML. Similarly, when academics and others discuss the dangers of AI, they are almost always talking about ML. It is important, therefore, to highlight some specific properties of ML that make implementations of ML successful and dangerous at the same time.

First, unlike GOFAI and Symbolic AI, which follow complex decision rules but do not evolve as a result of experience, ML is able to adapt ('learn'). During training (when ML algorithms are fed training data in order to 'teach' it something), the algorithm is able (through supervision or otherwise) to change itself when it gets things wrong or right. In a nutshell, if the ML algorithm gets something wrong, it adjusts so as not to make the same mistake again. After training, the ML algorithm may still adapt and learn while 'in the wild'. This can make the ML algorithm much more

[4] For a comprehensive list of terms and definitions see the Access Now report (8)

[5] For more details on the definition, main capabilities and scientific disciplines please see the EC HLEG full report available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (Retrieved on July 7, 2019).

robust when dealing with inputs that are constantly changing (e.g. an algorithm tasked with detecting cyber-attacks).

Second, ML does not rely upon rules programmed by human beings, as was traditionally done with symbolic AI. Symbolic AI could mimic human intelligence by combining a large set of rules into one algorithm. Depending upon the input, the algorithm would follow a path of rules leading to an appropriate decision. The greater the complexity, the greater the illusion of intelligence. ML provides an even greater illusion of intelligence in that it makes up its own rules and, as stated in the previous paragraph, can update those rules as it encounters new inputs.

Third, whereas a GOFAI or Symbolic AI systems could be clearly summarised with a list of decision trees and rules, ML algorithms are often opaque with regard to how decisions are made by the AI. That is, the 'rules' it makes for itself are not rules that we are often able to understand. While humans will look at features like 'two large ears, a snout, and a tail' to identify a dog, ML algorithms may use features and patterns that are not articulable in human language. Therefore, even if we could break through the opacity of the algorithm, we would not be able to understand what we saw. Thus, while ML has the potential to be an extremely powerful tool for high-impact decision-making, its inability to explain itself leaves us ignorant to the justification of those important decisions.

In summary, the aspect of learning has become central recently and is generally discussed in reference to ML systems. Describing AI as ML diverges from GOFAI in so far as an ML system is able to adapt its reasoning rules and decision making through an evaluation of its action, it: 'is a rational system that, after taking an action, evaluates the new state of the environment (through perception) to determine how successful its action was, and then adapts its reasoning rules and decision making methods' (29). It is significant to note that the human operator does not, in many instances, understand the reasons why one or another decision has been produced by the algorithm.

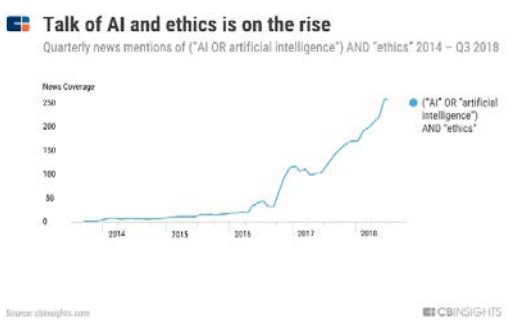
Of equal importance in a discussion of AI is the term 'algorithm'; AI/ML are specific types of algorithms. Algorithms are understood today as both a set of instructions and the execution of those instructions: 'Until we issue a command, or order an action, we have not conveyed an algorithm' (30). AI and ML are the names for a specific kind of algorithm. The difference between AI and ML as algorithms has to do with the rules governing the taking of a specific input and creating an output (see above).

To summarise, algorithms are a set of instructions combined with the execution of those instructions and AI and ML are different types of algorithms. GOFAI are given rules to categorise large amounts of data. In contrast, ML methodologies use algorithms that do not necessarily have rules at the outset, but develop or change rules as they interact with data. AI and ML algorithms have shown that these new kinds of algorithms present an extreme opacity in the work that the algorithm is doing.

3.1. Ethics of AI

In the last decades, academics have uncovered a range of ethical issues pertaining to AI (18, 31–35). Some of these issues relate to how AI/ML algorithms are made, e.g. how the data is acquired, sourced, and labelled (36); the computing power required to train an algorithm (37); the asymmetry in power, and the lack of transparency, between the private companies who have both the data and the computing power, and the consumer, who is reliant on private companies for their services (11). Relatedly, there are ethical issues resulting from how the AI/ML algorithm is applied in society, for example: facial recognition in public spaces as a threat to privacy in public spaces (38); differential impact in society seen through an unequal distribution of risks and benefits between groups (2,3); the potential for consumers to be unknowingly nudged to act in certain ways (15); lack of opportunity for meaningful, explicit informed consent (39); and the threat to constitutional democracy if AI/ML applications influence political power and the decision making of citizens (39).

Figure 1 – The rise of AI and ethics mentions in the media



Source: CBInsights

The last five years have seen a surge in talk about AI ethics (see Figure 1) in general with a particular emphasis on the role of 'black boxes' (11, 14, 40) and algorithmic fairness (3, 12, 19). Let us take a closer look at these two significant concerns, arising from the uniqueness of AI/ML as a technology, for a better understanding of their origin and meaning.

3.1.1. The 'black boxes' of AI

It is now commonplace to describe the workings of AI algorithms as 'black boxes,' the idea being that how a particular input (or inputs) results in a particular output (or outputs) is opaque. This term, however, conflates three distinct opacities – all of which are ethically salient.

The first sense in which AI can be a black box is when the technology is too complex for the average user to understand how it works. For example, autonomous cars rely on many sensors in order to drive. Consumers who do not understand the shortcomings of these sensors will be unable to make informed choices about when and where to engage the autonomous function of their car – like choosing not to use autonomous mode in the snow, where autonomous cars do not perform well (WeForum Study).[6] While these technical details may be well known to engineers, the media, the companies, etc. it may be unreasonable to expect the average consumer to be so informed.

The second sense in which AI can be a black box is when institutions intentionally obfuscate how their technology works. They may know perfectly well what considerations will be used to make a decision; however, they have reasons (good or bad) to keep the decision-making process a secret. For example, an intelligence agency employing AI to determine which people in airports to target for searches may not want to disclose the factors that trigger the AI algorithm to label you as suspicious. If malicious actors were to know these behaviours, then the technology would not be helpful. In other cases, companies simply want to protect their intellectual property – for example, Facebook may not want to divulge how its News Feed technology chooses which posts to show you. In this example, the obfuscation can be difficult to accept due to the negative consequences

[6] https://www.weforum.org/reports/reshaping-urban-mobility-with-autonomous-vehicles-lessons-from-the-city-of-boston

caused by these technologies – like the proliferation of fake news. Frank Pasquale has written extensively about this kind of algorithmic opacity in his book *The Black Box Society* (11).

The final sense in which AI can be a black box – and the sense that is truly unique to AI technologies – is that, in many cases, how the algorithm 'decides' upon a certain output is opaque to even the programmers who wrote the code (this is the case for the class of algorithms that fall under the umbrella of ML, see Section 4). This is often referred to as algorithmic opacity. Although we know that ML algorithms rely on statistical correlations between features of the input and the target, it is currently not possible to know what features the ML algorithm uses. Although much work has been done to try and figure out what these features are, there has been little progress. Google, for example, tried to reverse their picture matching algorithm to understand what the algorithm thinks things look like. You could search for 'bike' and the algorithm would output a dream-like image having wheels and handlebars.

The many black boxes of AI make it clear that there are many uncertainties when it comes to the functioning of the AI/ML algorithm, and yet there are at the same time known ethical issues as a consequence of this uncertainty. We will dive deeper into how these black boxes can be understood from the ethics and technology perspective later in section 5.

Insight #1

- Transparency of AI algorithms can mean three distinct things: first, the complexity of modern AI systems – leaving lay users in the dark; second, the intentional obfuscation by those designing AI solutions – leaving lay users and policy-makers in the dark; and third, the inexplicability regarding how a particular input or inputs result in a particular output or outputs – leaving everyone in the dark.

3.1.2. Biases of AI algorithms

The issue of algorithmic bias and/or algorithmic fairness has been intensely studied in the AI/ML debate (1, 6, 12, 41, 42). The reason for this comes from the fact that real world uses of AI have resulted in the unfair treatment of certain groups calling into question the fairness of the algorithm itself, i.e. the classification of data (e.g. the classification of certain groups as high vs low risk in predictive policing algorithms) and/or the fairness of the data used to train the algorithm[7] (e.g. the Amazon AI recruitment tool that showed preference for male candidates based on company training data).

Simply put, an algorithm requires data in order to learn patterns and/or generate rules about the data (this data is referred to as the training data). The goal is to have an accurate sampling of data, representative of the population, in order to train the algorithm accurately while also being as fair to different groups as possible. Decisions have to be made about how to classify items, animals, people or groups and how to qualify these classifications in order for the AI/ML to function. It is crucial to note that decisions about how to classify groups will, in some instances, be the result of cultural stereotypes and prejudices. Algorithms are already being used for pre-trial decisions about whether or not defendants should be released back into the community or should remain in jail and 'in some cases, black defendants are substantially more likely than white defendants to be incorrectly classified as high risk' (1).

AI algorithms have also been used in the hiring at companies such as Amazon, only to show that preferential treatment was given to male applicants who were predicted by the algorithm to be

[7] See also the report by the Big Brother Watch group in the UK discussing the problems of the training data for predictive policing algorithms in the UK and the resulting biased and discriminatory decisions, https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf (Retrieved July 14, 2019).

more desirable/successful than female applicants.[8] In both of these instances, a problem arose concerning the fairness of the algorithm based on either the way in which data was classified or the training data that was used for the algorithm: the data and/or its classification exhibited cultural biases and stereotypes that were exacerbated in the resulting algorithm, i.e. that black defendants are high risk versus white ones or that men perform better in high-level jobs than women. The result of these biases was a differential impact on one group of persons over another.

To be sure, we are talking about the use of AI/ML algorithms in diverse contexts where the potential impact of a biased model on individuals could be catastrophic, including social services (2), predictive policing for allocation of police throughout the city (3) and other public services. The issue at hand here is that historical and governmental data is used to train the algorithm, and this data is the result of cultural biases and stereotypes. Judges, for instance, have historically labelled black defendants as high risk and this bias is a starting point for training an AI. In the case of AI for child welfare, 'some communities—such as those in poverty or from particular racial and ethnic groups—will be disadvantaged by the reliance on government administrative data' (2). For policy-makers to be able to protect vulnerable demographics, it is paramount to understand 'the downstream consequences of AI' (12) and the value trade-offs discussed in the design process so as to open such trade-offs for critique.

The problem raised of late is how to be fair to different groups – what does 'fair' mean (is it about data sets being accurate or about containing every possible bit of data)? What does 'accurate' mean (who determines if data is accurate), and in which instances is it better to be accurate even at the cost of fairness or vice versa, better to be fair even at the cost of accuracy? 'To address these issues, practitioners will sometimes be forced to make value trade-offs between competing and incompatible notions of bias or between human versus machine bias' (12). It is these questions precisely that call for contributions from the field of ethics to: clarify meanings, uncover ethical issues, and imagine possible solutions.

Insight #2

- Bias and fairness of AI/ML algorithms resulting from the training data is a significant barrier to the ethical development and use of AI/ML. Important questions concerning what is 'fair', what is 'accurate' and how to balance trade-offs between the two must be analysed.

[8] For more see https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (retrieved July 9, 2019).

4. Ethics

The last three years have seen an unprecedented number of AI ethics principles developed by governments, civil society, private companies, and multi-stakeholder groups. The 'Principled Artificial Intelligence Project from the Harvard Law school[9] has charted thirty-two of these initiatives from 2016 until mid-2019 (e.g. among them are the European Commission High-Level Expert Group on AI, the Montreal Declaration and many others). These sets of principles are a response to the growing number of AI/ML applications and the increased awareness of ethical issues resulting from these applications (e.g. differential impact in society etc.). Further, they are a suggestion about how AI should be built and used. The idea being to capture the moral dimension of AI.

For decades, the field of 'ethics of technology' has been producing insights regarding the ethical issues of modern technologies. AI is a technology, and therefore, it will be important to go over the major insights from the ethics of technology before tackling the specific issues to AI. First, however, it is important to say a word or two about what 'ethics' is.

4.1. Ethics

The study of ethics, as a branch of philosophy, is centuries old. It concerns itself with questions such as 'what is a good person', 'what is a good act', 'what is a good life' and 'how does one achieve the good life?'. It is a field tasked with characterising distinctions between good and bad, and right and wrong. It is directed at addressing questions concerning the good life for both the individual and the community. As a discipline, it fosters balance between competing interests alongside competing conceptions of the good life through constant critique and reflection of things often taken for granted. More specifically, it is about: 'promoting objective (but context & culture-dependent) conditions of human flourishing; respecting the dignity of others and the duties created in our relationships to them; living as a person of integrity and principle; promoting beneficial and just outcomes while avoiding and minimising harm to others; cultivating one's own character to become increasingly more noble and excellent; the skilful practice of moral perception, sensitivity, and discerning judgment; and learning to more expertly see and navigate the moral world and its features'.[10]

Ethics and/or ethical deliberation is an ongoing process, much like character development, directed at understanding what it means to be a good person and to live a good life: 'ethics cannot be approached like mathematics; there is no algorithm for ethics, and moral life is not a well-defined, closed problem for which one could design a single, optimal solution. It is an endless task of skilfully navigating a messy, open-ended, constantly shifting social landscape in which we must find ways to maintain and support human flourishing with others, and in which novel circumstances and contexts are always emerging that call upon us to adapt our existing ethical heuristics, or invent new, bespoke ones on the spot'.[11]

A task for the ethicist when discussing ethical issues at large is to describe and make clear what the ethical issue is and how it comes to be understood as such. It is the task of ethics to engage in the search for, and articulation of, problems as much as it is the task of ethics to look for solutions to said problems. What's more, it is the task of ethics to call attention to the variety of problems and

[9] See https://clinic.cyber.harvard.edu/2019/06/07/introducing-the-principled-artificial-intelligence-project/ (Retrieved on July 13, 2019).

[10] The Marrkula Center for Applied Ethics, a part of Santa Clara University, has created a program for Ethics in Technology Practice to assist in the implementation of ethical reflection within the corporate technology space. For more on this, see https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/ (Retrieved Feb 3, 2019)

[11] From https://www.scu.edu/ethics-in-technology-practice/conceptual-frameworks/

solutions as a way to empower decision making on the part of citizens, politicians, technologists, educators, policy-makers, etc.

Seen through this lens, it must be acknowledged that searching for answers in ethics will not resemble the kinds of solutions one encounters in the natural sciences – there may not be a well-defined problem to solve let alone a single universally accepted solution to said problem(s). Furthermore, any set of guidelines, codes of conduct, principles, or laws will inevitably fall short of capturing ethics in its entirety. When general concepts are made concrete, some piece of the broader sentiment can be lost in implementation. Moreover, it may not always be possible to create technical solutions for ethical problems. Consider for example poverty, AI tools can help do things to make legal counsel, financial guidance, and job searching more accessible to low income populations; it might help provide treatment for substance use or post-traumatic stress disorder which make it harder for people to emerge from poverty; but it cannot single-handedly solve the systemic issues that make poverty happen.

Ethics is also about understanding concepts such as 'moral overload' – in essence, that 'we are repeatedly in situations in which we cannot satisfy all the things that are morally required of us. Sometimes our moral principles and value commitments can simply not all be satisfied at the same time given the way the world is' (43). Situations such as these are also referred to as moral dilemmas or conflicting preferences, and demand that we address the various options available to an individual to act and/or to deal with the 'moral residue' after the fact, i.e. the 'moral emotions and psychological tensions that are associated with the things that were not done, the road not travelled, the moral option forgone' (43).

Ethical theories, also known as theoretical frameworks, 'help you recognise ethical issues when you are in their presence, and help you to describe them.'[12] Accordingly, ethical theories work as a 'field guide' to help not only identify ethical issues but to help describe and eventually overcome them. Three of the more mainstream ethical theories currently in debate (but these are by no means the extent of the ethical theories available) are: deontology, consequentialism, and virtue ethics.

Deontological ethical frameworks 'focus on moral rules, rights, principles, and duties'. These rules and principles are thought to apply to all cases, which often results in a conflict of prioritisation between one of more rules or principles. Consider once again the bioethical principles of autonomy, beneficence, non-maleficence, and justice (16). In the case where a person is in need of treatment but receiving such treatment would conflict with his/her religious beliefs the physician is confronted with a conflict between the duty to do good on the one hand and the duty to respect the self determination of the patient on the other. When such conflicts arise, the challenge is to identify the duties carrying the most ethical weight in a given situation[13]. Some of the main duties or principles that are central to a deontological framework are: Rights of individuals and groups (i.e. people are entitled to certain economic, civil, religious, and moral protections); autonomy (i.e. that individuals should be free to choose for themselves); fairness (i.e. the requirement to equally distribute goods, wealth, harms, and risks); and universality or consistency (i.e. that all persons should be held accountable to the same standards). There are, however, a few difficulties with deontological frameworks: how does one know unequivocally that one duty ought to be prioritised above another? How long is the list of duties, and how does one prioritise duties when they come into conflict with one another? Consider the patient above who refuses lifesaving treatment on religious grounds, if the conflict is between doing good and respecting autonomy, how must the doctor decide which one is more important? In many instances additional information is needed, for example is the patient a child or an adult, is there any other option for treatment, does the patient

[12] See https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/

[13] Other prominent deontological frameworks include the work of W.D. Ross (1877-1971) and of Immanuel Kant (1724-1804). This discussion is not exhaustive by any means but is meant as an introduction to what ethics teaches us.

know their life will terminate if they do not accept the treatment etc., and yet still there is no easy answer to this.

Consequentialist ethical frameworks, in contrast to deontological ones, determine that an action is good or bad by the consequences of the action. Although there are many instances in which the scope or breadth of consequences are not known, there are other cases in which the consequences can be determined and as such 'can be readily seen as morally choice worthy ('this is the best thing for us to do'), morally permissible ('it's not wrong for us to do this'), or morally impermissible ('we shouldn't do this, it's wrong'). When the moral consequences of a technological choice are sufficiently foreseeable, we have an ethical responsibility to consider them'.[14] There are, however, a few difficulties with consequentialism: it provides no guidance on how to address the unintended consequences, on how to predict all the consequences, on which consequences matter, and on whose consequences matter, etc.

One of the more common forms of consequentialism is utilitarianism[15]. Utilitarianism asks us to weigh the overall happiness (measured as the aggregate pleasure and the absence of pain) or welfare that an action can bring about for all those affected. A difficulty here is to properly identify *all* the possible outcomes for *all* the possible stakeholders affected. Still, this conceptual framework is appealing for engineers because it presents the possibility to quantify and calculate the ethical variables of an action. As Vallor et al. point out, however, 'this is often an intractable or 'wicked' calculation, since the effects of a technology tend to spread out indefinitely in time (should we never have invented the gasoline engine, or plastic, given the now devastating consequences of these technologies for the planetary environment and its inhabitants?); and across populations (will the invention of social media platforms turn out to be a net positive or negative for humanity, once we take into account all future generations and all the users around the globe yet to experience its consequences?)'.[16]

The third main ethical theory often discussed is virtue ethics, which states that an action is ethical if one acts according to what a virtuous person would do. A virtuous person is conceptualised as an individual who exhibits virtues (aka character traits) that are necessary to be a good individual – one who can flourish in life. Virtues such as courage, honesty, and integrity, steer one to act in a good way. Virtues must be the perfect middle point between two extremes; too much of a virtue (e.g. too much courage to the point where one is considered foolhardy) or not enough (e.g. not enough to the point where one is considered a coward). The difficulty with virtue ethics is that it is often quite difficult to articulate which virtue is superior to another for a given action or situation. Moreover, can a person be considered virtuous if the net consequences of his/her action was negative?

Each of these ethical theories has its strengths for identifying what the good or the right thing to do is and each of these ethical theories has its downsides. The benefit of knowing about these theories is that people (e.g. technologists) often employ them, even implicitly, in their daily lives, or in the creation of a technology. By making use of these conceptual frameworks explicitly we can help critique and refine the decision-making process. For example, if one wishes to claim something is good or ethical based on the consequences of the technology, then one is working from a consequentialist framework and should also consider the difficulty in predicting and/or expecting all the possible consequences and for whom. If, on the other hand, one wishes to present principles or duties for ethical grounding, then one is working from a deontological framework and must also be prepared to account for the strain when principles come into tension with one another and the difficulty in reasoning through which principle is more important and why.

[14] See https://www.scu.edu/ethics-in-technology-practice/conceptual-frameworks/

[15] Formulated by John Stuart Mill in the 1800s, see (44)

[16] See https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

4.1.1. Key lessons

In order to capture ethics as a resource, one must understand that ethics provides a variety of conceptual tools for understanding how to evaluate actions and people. Some of these ethical tools may be partially translated into technical solutions whereas other tools (e.g. dealing with moral overload) cannot be. What ethics as a study provides us with is the capacity, and tools, for deliberation about the kinds of people we want to be, the kinds of communities we want to build, and the kinds of technologies we want to create and use. The goal of ethics is to provide strong enough rationale that an individual is compelled to act in a way they believe is the right/good way. The lessons we can take away from understanding ethics in this way are:

- Ethics cannot be reduced to codes of conduct, guidelines, or principles exclusively. Rather, ethics should also be understood as a continuous process (akin to character development) that must accompany the design, development, and implementation of AI;
- Engineers may be confronted with moral overload in the AI development context and there needs to be a system in place to deal with these aspects for technologists, implementers, and users;
- There are a variety of ethical theories, aka conceptual frameworks, that AI developers will implicitly use in decision making and it is important to be able to identify which conceptual framework is being used and to know the benefits and risks of each of these.

4.2. Ethics and technology

One will be hard pressed to discuss today's concept of the good life, good actions, or good people without being confronted with technology. Technology is pervasive in today's world (June 2019) and it therefore contributes to our decision making, value systems, assumptions, biases, and experiences of daily life. The field of 'ethics and technology' is tasked with asking how technology impacts the good life of people, our ability to be good people, and to act well. In fact, 'in the evaluation of new technologies a host of moral questions are relevant, some about risks, some about how we are to design new technologies, some about how we are to distribute responsibilities for new technologies, some about the proper role of the government etc.' (45). Some of these questions are directed at the makers and regulators of technology, while others are directed at the users of technology. Some questions even concern the moral status of the technology itself: should algorithms or robots have rights which protect them, or be expected to have moral reasoning capabilities?[17]

Although reflections on technology date back to Ancient Greece, the most oft cited philosopher of technology for many is Martin Heidegger, who questioned the inherent power of technology to influence and steer human behaviour. More recently, the philosopher of technology Shannon Vallor has ushered virtue ethics into a discussion of how technology can be used to assist in human flourishing and character development as opposed to the more dystopian 'technological determinism' of earlier years (47). Ethics and technology provide a host of insights into the relationship between humans and technology. When we understand the questions being asked from this space and the themes raised, we can then appreciate the need for AI ethics guidelines or principles and the governance thereof. The specific points of attention from the field of ethics and technology that we will call attention to are, that: AI should be understood as a socio-technical system; the risk in believing technology happens without society having control over it (also known as technological determinism); there is a need to capture the qualitative risks and uncertainties associated with AI innovation; and, AI technologies used in the real world should be framed as a

[17] For more on this discussion see (46)

'social experiment'. The understanding of these specific points will provide us with the evaluative tools necessary to craft appropriate legislation and policy regarding AI technologies.

4.2.1. AI as a socio-technical system

In contrasting technology from science, philosophers of technology have argued that technology and science are differentiated from one another in so far as science concerns itself with what is (and the study of what is), whereas technology concerns itself with what is to be, or what ought to be and the creation of technologies to enforce how the world ought to be.

Political philosopher Langdon Winner (48) asked whether technology has politics, or better put, whether technology carries human biases. In his analysis of bridges built in New York City from roughly the 1930s to the 1960s, he claimed that the bridges built by Robert Moses were done in such a way that they manifest a specific prejudice of Moses. Effectively, the bridges were made in a way that prohibited African Americans from going to the wealthy beaches, leaving only the affluent white populations to attend. The bridge, in this case, was a physical manifestation of technology that reinforced a common societal prejudice at the time. Of equal importance, the bridge made a statement about how the world should be – in this case segregated. It is important to note here that while Winner gives an example of explicit prejudice, many prejudices are often implicit. This does not mean that designers are necessarily malicious, but personal biases or prejudices may be unknown to the technologist. In many AI applications to date, societal stereotypes are exacerbated through the training and use of the AI system. Data acquired from society, which contain societal biases, are used to train the system, and thus the biases are also trained *into* the system. Just as the bridges in New York were built to physically enforce cultural prejudices, AI has been used to digitally enforce societal and cultural prejudices in law enforcement (e.g. predicting that African American male defendants have a higher risk of re-offending) and job recruitment (e.g. that men are better suited to high-level professional positions over women).[18]

In the field of Science and Technology studies (STS), Madelaine Akrich discusses the assumptions designers make regarding the distribution of roles and responsibilities of actors (including human, environment, and technical objects as actors). For Akrich, 'many of the choices made by designers can be seen as decisions about what should be delegated to a machine and what should be left to the initiative of human actors' (49). By making choices about what should and should not be delegated to certain actors (human or nonhuman), engineers may change the distribution of responsibilities in a network. In certain AI applications, for example the Dutch grocery store chain Albert Heijn, we can already see that choices are being made about the kind(s) of tasks that humans should do versus the kinds of tasks that machines should do. Albert Heijn is using an AL algorithm to help reduce food wastage by automatically discounting chicken and fish products based on their sell-by-date 'with the higher discount for items that need to be sold soonest',[19] whereas humans are stocking shelves and delivering groceries. In short, technology developers are not solely creating things to reinforce the way the world works (in a descriptive way), they are also involved in prescribing the way the world should be, with either/both positive and negative outcomes.

This prescriptive role of the designer/technologist brings us to a discussion of how values (and not only biases or assumptions about roles and responsibilities) are embedded into technologies. Ethicists and philosophers of technology began the discussion of whether technology is value neutral or value-laden back in the 20th century. According to the neutrality thesis, such systems are in themselves neutral and depend on the user for acquiring a moral status as either good or bad (think: 'guns don't kill people, people kill people'). In contrast, the embedded values position argues that it is possible to identify tendencies within a computer system or software that promote or

[18] See https://www.cnbc.com/2018/10/10/amazon-scraps-a-secret-ai-recruiting-tool-that-showed-bias-against-women.html (Retrieved July 13, 2019).

[19] See https://www.salesforce.com/company/news-press/stories/2018/12/121018-i/ (Retrieved July 13, 2019).

demote particular moral values and norms (50, 51). These tendencies manifest themselves through the consequences of using the object (e.g. cookies on websites facilitate tracking of an internet user's moves ultimately demoting the value of privacy online). When a technology is capable of imposing a behaviour on a user, or there is a specific consequence to using the technology, the imposing force within the technology that steers this is considered a 'built-in' or 'embedded' value (or alternatively a disvalue if the computer system hinders the promotion of a value) (50–53).

Given that technologists can play a prescriptive role without conscious intent the question arises of how to account for this tendency. One possible role of the ethicist can be to make such assumptions, biases, and/or values explicit and open to critique through dialogue between ethicist and technologist (54). Ethical theories for the approach to design have taken this even further to suggest that designers ought to make their value conceptions, interpretations, and tensions explicit and open to scrutiny as part of a systematic, democratic and ethically sound design process (52, 54–57).

For many scholars today it is well known that in design, engineers are working to promote values in their designs – values such as safety, security, privacy, human welfare and of late sustainability. At the same time, it is also well known that 'when we design not for one value but for a range of values, it will regularly occur that design options that score good on one value score less on another' (52). Consequently, engineers are left with value conflicts and must make choices between one or the other. The question then becomes: how does one make such a choice? Of equal importance in such a discussion is that values are not static but are changing (24), e.g. the value of privacy in the medical space used to refer predominantly to one's physical/corporeal body while today privacy is most often attributed to control over one's medical data.

The significance of discussing the reciprocal relationship between society and the technologists creating value-laden AI is also to point out that we cannot ethically assess AI/ML if we envision society (and the places where AI/ML will be applied) separate from the making of AI/ML and the many stakeholders involved in this (e.g. technologists, financers, activists etc.). The social and the technical are deeply connected, as shown in this section. Moreover, the technical does not impact individuals alone but entire infrastructures and systems are impacted upon the introduction of certain technologies (e.g. AI/ML impacts the healthcare system rather than a lone hospital). A consequence of this is that we must envision AI/ML as a socio-technical system (58, 59): AI/ML should be understood as a complex confluence of both society (people) *and* technology, rather than society and technology isolated from one another until the moment AI/ML is introduced into the real world. The consequence of this is that AI/ML should be evaluated with reference to the society in which it has been created. Further, that society's role in the development and applications of AI/ML should not be underestimated.

Insight #3

- AI should be understood as a socio-technical system and should be assessed according to the society in which it has been created, further, society's role in the development and applications of AI/ML should not be underestimated.

4.2.2. Technological determinism and the responsibility gap

The belief that technology determines the development of social structures and cultural values is widely known as technological determinism. Deterministic thinking leads one to assume that when we create AI/ML, the capabilities of the technology will dictate how we structure values and concepts rather than using societal values as the guiding feature to govern technology. This creates an influential rhetoric in the creation, or lack thereof, of policy and regulation for AI/ML.

Let us take the notion of responsibility as an example to address the power of deterministic thinking in the framing of technology and its regulation. The lack of transparency and/or control that humans have over the generation of rules in a ML algorithm has introduced the notion of a 'responsibility

gap' (4). For some, 'because certain artificial agents learn as they operate, those who designed those agents may not be able to control or even predict what their agents will do. As these agents become increasingly more autonomous, the argument goes, no humans will be responsible for their behaviour' (60). Consequently, there is a responsibility gap insofar as no human is responsible for the consequences of the AI/ML. In response to such claims – that the control requirement must be met in order to assign responsibility to a human – some have argued that in other contexts we 'use a variety of conceptual frameworks and technical tools … which enable one to deal with problems of responsibility ascription without appealing to the control requirement' (61). The point being stressed here is that 'there are situations in which we hold humans responsible for outcomes that they could not control. Strict liability is an obvious example here' (4). For others, professional responsibility provides the grounds for holding human engineers responsible for the behaviour of artificial agents (62). Each of these perspectives aims at countering the tendency to assume that because the technology is a certain way now (unpredictable), this determines who is and is not responsible for the outcomes it produces.

In contrast to this deterministic view (i.e. that the technology controls the attribution of responsibility), other scholars insist society must be reminded that 'whether or not there will ever be a responsibility gap depends on human choices, not on technological complexity' (4). The complexity of technology and the accountability relationships formed around a technology are not rigidly/permanently defined – rather, they are socially constructed. Further, 'in order to imagine a future time at which there will be artificial agents for which no humans are responsible, we have to imagine that the human actors involved would decide to create, release, and accept technologies that are incomprehensible and out of control of humans. In addition we have to imagine that the humans involved (especially consumers, users, and the public) would accept an arrangement in which no humans would be considered responsible for these technologies' (4). Even if technology is considered wholly 'responsible' one day, this would be allowed only through human abdication of control.

To be sure, responsibility is a complicated concept made even more so by the 'problem of many hands' – that many actors are involved in the development, production, and use of a technology and the various aspects of its operation (63). One view holds that notions of responsibility can be understood in terms of norms about accountability. In other words, accountability is embedded in relationships between those who have an obligation to a community (e.g. members of the public, consumers, disadvantaged groups etc.) and members of that community who believe they are owed an explanation if something goes wrong (i.e. what happened and how it happened) (4). It follows from this that just because ML is complex and uncontrollable (to a certain degree), this qualifier does not demand that society accepts 'an arrangement in which no humans are responsible for the behaviour of those agents' (4); rather, it demands that the variety of actors involved in AI/ML development, implementation, use, and regulation decide together with users about the accountability-responsibility relations they wish to enforce. In other words, 'responsibility requires that technology be designed and built so as to facilitate, or even ensure, human responsibility' (4).

Insight #4

- AI can be designed and implemented in ways that obfuscate attributions of responsibility and accountability, but that does not necessarily mean that responsibility and accountability are not possible in the context of AI.

4.2.3. Uncertainties and risk

With any new or emerging technology, and even more so in the case of AI/ML, there are uncertainties about the positives and negatives that it will bring. The term 'uncertainty' has a broad meaning but essentially refers to 'everything that we might wish to know, but yet do not know' (64). Uncertainty is a major complicating factor when creating policy for emerging technologies, as it is

difficult to understand and quantify the potential risks and benefits before the technology is in use, for example, one could not have anticipated the variety or criticality of long-term privacy risks when Facebook was first introduced. Making it even worse, the deeper you dive into a new technology to uncover risks, the more uncertainties you will encounter (64). For example, the more we learn about the carbon footprint associated with training one ML algorithm [20] the more uncertain it becomes what the long-term impact of AL/ML on society will be and who will bear the costs of the environmental consequences. While it could be argued that all uncertainties should be taken into account in a decision making situation, this will also make decision making 'extremely complex and time-consuming, thereby leading to delays and stalemates and in some cases possibly render us unable to make any decision at all' (64). Such stalemates within the context of AI are undesirable, and thus it is important to understand that while uncertainties need to be evaluated and prioritised, their existence is inevitable and should not prohibit regulation/policy.

Some of the possible approaches for evaluating and/or prioritising uncertainties come in the form of a risk assessment and/or risk analysis and include 'risk-cost-benefit analyses, and risk standards. One major problem with risk assessment and cost-benefit analyses is the 'strong tradition of quantification': 'The aim is usually to produce a quantitative assessment, and therefore the focus is on quantifiable factors, such as the expected number of deaths and the expected economic gains or losses. Values that are difficult or impossible to quantify tend to fall outside of such comparisons, for example cultural impoverishment, social isolation, and increased tensions between social strata' (64). In an AI context, many of the ethical considerations raised to date are of a qualitative nature: 'how we perceive and understand our environments and interact with them and each other is increasingly mediated by algorithms' (19). It is, therefore, paramount that tools for evaluating the presence and impact of such qualitative risks are developed.

Two other problems associated with risk analyses include 'the fallacy of undetectable effects' and 'the fallacy of disregarding benefits'.[21] The fallacy of undetectable effects, in short, is concerned with the idea that there 'may be strong reasons to believe that an effect exists even though we cannot discover it directly'. For example, the effects of a new technology on an entire population may not be detectable on an individual level, but from a public level, they are quite serious. In an AI context, for example, one may suggest that if we cannot observe negative results of biases in training data, because only a small (or large but marginalised) percentage are impacted then this imperceivable impact should not block the use of AL in predictive policing, detection of child maltreatment or other applications. However, it is important to remember that cases like the Amazon recruitment AI algorithm teach us how cultural biases embedded in historical data are exacerbated in the resulting technology and will be most noticeable when looking to the group rather than the individual (remember: females were favoured over males by the recruitment AI based on ten years' worth of Amazon practices used to train the algorithm). Thus, evaluating the AI/ML application must not focus solely on the impacts to individuals but those impacts on the group as well.

A related concept, the fallacy of disregarding benefits. One version of this fallacy is important to note and 'consists in using the benefits a certain risk provides in one context as an argument for accepting the same risk in contexts where these benefits do not arise' (64). The oft cited example of this fallacy concerns doses of radiation: 'Nuclear technology cures countless cancer patients everyday – and a radiation dose given for radiotherapy is no different in principle to a similar dose received in the environment' (66). Yet, 'This is a serious fallacy, since in oncology, the only chance to save the patient's life may sometimes be a therapy including high doses of ionizing radiation that

[20] See https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/ (retrieved July 13, 2019).

[21] 'The notion of a fallacy is not entirely clear. The Oxford English Dictionary uses the phrase "deceptive or misleading argument"' in defining it. This could be improved by observing that fallacies (in the philosophical sense) are argument patterns, rather than single arguments. We can at least provisionally define a fallacy as a "deceptive or misleading argument pattern"'' (65)

significantly increase the patient's risk of contracting a new cancer at a later point in time. Extensive epidemiological studies show that high dose radiotherapy leads to significant risks of new, radiation-induced tumors' (64). In a banking context where ML/AI is used, one might suggest that the risk to privacy (i.e. lack of control over one's data) associated with the collection and use of personal data comes with the benefit of fraud detection and, therefore, that the same risks to privacy should be accepted in a mortgage or loan application process. The same benefit of fraud detection (a benefit to both the bank and the client) does not exist in the use of ML for mortgage or loan processes and therefore the risk to privacy (as seen through surveillance type collection and use of data) may not be deemed acceptable. To avoid this fallacy, risk assessments must be appropriately calculated for each ML application.

The idea of risk assessment and cost-benefit analysis, however, has been criticised for a variety of reasons, namely 'they deny uncertainty and ignorance; second they short-circuit the moral dimension of new technological developments; third, they do not address the need for profound (social) learning from, for example, errors and catastrophes' (67). The first and third criticisms will be discussed in section 5.25 but the second criticism – that risk assessment short-circuits the moral dimension of new technological developments – requires further analysis as the purpose of this study is to grapple with how ethical analyses should be conducted in the context of AI.

Insight #5

- Risk assessments, while valuable, do not capture important ethical risks which may be unquantifiable, qualitative, or unobservable.

4.2.4. Ethical Technology Assessment

We are now confronted with a need for a form of risk assessment that goes beyond the traditional cost-benefit analysis approach – one that captures the moral dimension of AI above and beyond a quantitative risk assessment. One possibility for this is the 'ethical Technology Assessment' (eTA), which is an assessment that creates space to discuss the ethical issues related to, and resulting from, technology (and in this case AI) in a pre-emptive manner, i.e. prior to its deployment in society.

The idea behind an eTA is to have ethics play a more critical role in the assessment of a new technology. Further, to have ethicists involved in the process of assessing the technology insofar as their skills of ethical reasoning are put to use in the evaluation of the technology by 'recasting the way problems are defined, by exploring the interrelationship of the technical and non-technical issues, and by analysing technology itself as problematic' (68). The eTA is thus a starting point for the ethicist to engage in a critical reflection about a specified scope of topics and/or issues.

eTA is presented in the form of a checklist. Any version of an ethics checklist will be criticised insofar as ethics cannot be reduced to a checklist (see section 5.1); however, the intention with the original eTA (provided by the scholars arguing in favour of an eTA) is to show a vision of the most common ethical issues (identified through historical experience) and to make these the starting point for discussing the ethical acceptability of a technology. The checklist is not meant as a one-off assessment; rather, the aim is to stay in contact with the technology developers and to have continuous dialogue throughout the life cycle of the technology using the criteria of the checklist as the focal point.

The following criteria have been put forth in the original version of the eTA: '1. Dissemination and use of information; 2. Control, influence and power; 3. Impact on social contact patterns; 4. Privacy; 5. Sustainability; 6. Human reproduction; 7. Gender, minorities and justice; 8. International relations; 9. Impact on human values' (65). Using these nine criteria as a guideline the ethicist can then work together with the technologists to form an assessment on how the technology in question will impact society according to these common themes.

In order to tailor the eTA to avoid abstract ethical discussions, one can also make an eTA more concrete by: specifying the context in which the AI will be used (e.g. policing, education, etc); the AI methodology used (e.g. GOFAI, ML, etc.); and the stakeholders involved (e.g. citizens, police officers, migrants, etc.). Including each of these specifications may help to understand how societal values are interpreted and prioritised as this may differ from one context to another or from one stakeholder group to another (21). Consider, for example that privacy in the healthcare space refers to medical data, and in some instances, the physical body (i.e. things that happen in healthcare), while privacy in the home may refer to familial details like parental relationships, shopping habits, or music preferences (i.e. things that happen in the home). Specifying the context, stakeholders (or actors) and AI methodology can help create a more grounded and fine-grained picture of the ethical concerns.

With this in mind, the ethicist can begin an analysis of possible ethical concerns related to a particular AI/ML application. To give a short demonstration of the kind of ethical reflection that might be found in an eTA, let us take the application of predictive policing in a neighbourhood in which there are many migrants living, and in which there has historically been a high level of criminal activity. Let us imagine that the idea has been put forward to use an AI algorithm to predict potential criminal activity before it happens. In an eTA for a case like this, the ethicist (working together with the technology implementer or developer) would ask questions about the motivations for using AI in the first place and would use the nine criteria listed above (65) to structure a systematic reflection on the specific prospective (or ongoing) application. Rather than go through each of the nine criteria here, let us take a look at just one of the criteria, #3 'Impact on social contact patterns'. The ethicist will immediately raise a number of concerns. The first may be concerns for police officer assumptions of innocence changing; the prediction of the algorithm may confuse the possibility of a crime happening with the likelihood that the crime will happen (i.e. that the crime is *going* to happen). This could create a feeling of insecurity on the part of people who happen to reside in an area where there is historical data showing an elevated occurrence of crime or where a history of a certain demographic points to a likelihood for criminal activity.

Adding weight to this line of thinking, a United Nations report raised concerns over predictive policing tools that they could 'change how law enforcement sees the communities they patrol and influence important decisions such as whether to make arrests or use force. Bias may also lead to the over-policing of certain communities, heightening tensions, or, conversely, the under-policing of communities that may actually need law enforcement intervention but do not feel comfortable in alerting the police' (69).

Given the lack of effectiveness shown for predictive policing tools to date, the lack of transparency with their use, and the historical bias in that sector, there exists a serious risk that patterns of contact between police officers and community members will change from each being members of the same community to each being antagonistic with one another. Moreover, the threat to civil liberties, such as freedom of movement and the presumption of innocence are in jeopardy through the use of AI for prediction of crime in neighbourhoods. And this says nothing of the risks to data privacy, considering the kinds of data that will be collected in order to run the algorithm, and the sensitivity of the data that the algorithm will generate. More could be said on the additional suggested applications of AI in predictive policing, e.g. to identify suspects in public spaces through facial recognition, to assess individual's risks of committing crimes, and supporting police investigations in general.[22] If the intention for using the algorithm in the first place is to reduce crime, one might suggest the alternative approach of community policing as a more promising solution (70).

[22] For more on this see https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf (Retrieved July 14, 2019).

To summarise, an eTA for a case like this should be clear about the proposed AI application (i.e. an AI application within the police force used to proactively predict neighbourhoods where crime might take place based on historical data of crime throughout the city), the range of stakeholders who will be involved and/or impacted, and an analysis of the nine criteria of ethical concern. The goal in doing such an eTA is to make policy options for limitations of the application and/or safeguards that ought to be put in place.

With one of the threat's being a transformation in contact between police officer and individuals in the neighbourhood, one might suggest that all neighbourhoods be assigned consistent officers that maintain a strong connection with the neighbourhood and in the event that an AI/ML algorithm predicts crime, additional police officers are sent to the area to work together with the officers already in place. Additionally, if innocent citizens will be subjects in such an AI trial one would suggest they be given the opportunity for informed consent and made aware (at the very least) of the experiment about to take place through town hall meetings and the like. Such an action reaffirms the responsibility of public service officials to be transparent about their means, and the limits to these means, of policing.

Another safeguard may concern the data generated by the AI algorithm; predictions about a neighbourhood being more prone to crime could have a negative impact if made available outside the policing context (not to mention inside the policing context). Such data could be used to: deter people from visiting the neighbourhood (through an online map or otherwise), impact housing prices, profile individuals applying for loans and/or admission into academic institutions. Consequently, this data should be treated as highly confidential within the police force and prevented from any kind of secondary use.

Relatedly, other safeguards regarding this application may be that only one neighbourhood at a time should be subject to this kind of predictive policing and furthermore that a time limit be put in place to ensure that the application is used for a short period of time after which an evaluation of success (with specified metrics) must be completed. Given the risk for alienating the community, members of the community should be interviewed as part of an evaluation of the technology.

It is also important to discuss the limitation of the proposed application, thus the eTA is also meant as an opportunity to prevent the inclination towards 'AI solutionism', i.e. the idea that AI will be thought of as the answer to all problems no matter the scope (a version of technological solutionism as introduced in Morozov (71). There will be, in many instances, other alternatives to the identified problem besides the use of AI/ML. For the above example of predictive policing, if the aim is to deter crime then AI is not the only solution available, the practice of community policing also shares the same goals with a history of success.

The above reflection was an example of the kind of content one might find in an eTA directed at AI. A reflection on the kinds of ethical concerns raised by the application with insights into the possible safeguards and limitations that may accompany the use of the AI. Given the earlier discussion - that AI/ML should be understood as part of a socio-technical system, that technology need not determine the conditions of human responsibility (and/or accountability), and that AI is ushering in new ethical issues resulting from its design and deployment - an eTA may be a tool to capture these themes and make them tangible. In particular, the use of an eTA before the AI/ML is applied can help facilitate the chance for dealing with ethical problems faster and at an earlier stage (65).

Insight #6

- Ethical technology assessments are a viable mechanism for uncovering novel ethical issues which may arise due to the development and use of AI.

4.2.5. Technology as a social experiment

Adding further complexity to the ethical discussion is that the AI socio-technical system is a dynamic one rather than a static one. The values of concern may change over time (e.g. the meaning of privacy has changed with the availability of personal data online see (72); the risks and uncertainties associated with AI/ML will change (i.e. today there is an emphasis on the risks of algorithmic bias but in the years to come the environmental risks may increase in criticality); the capabilities of AI/ML solutions may change (i.e. from assistance in decision making to autonomous decision making); and the range of applications will undoubtedly expand. This dynamic nature of both the social and the technical structures at play demand that we frame the AI socio-technical system accordingly, as evolving rather than as set in stone.

One possibility to capture this dynamism is through an understanding of technology as a social experiment. The 'technology as a social experiment' approach considers technology experimental 'if there is only limited operational experience with them, so that social benefits and risks cannot, or at least not straightforwardly, be assessed on basis of experience' (67). Naturally, there will be varying degrees of operational experience with the technology depending on the length of time and the kind of experience one hopes to gain; 'there is now more than fifty years of operational experience with nuclear energy which makes this technology no longer experimental in some respects, but this experience is arguably still very minor when it comes to the issue of nuclear waste disposal, which is to be stored safely for periods up to 10,000 years' (67).

Framing AI as a social, or real world, experiment encourages 'the gradual and experimental introduction of a technology into society, in such a way that emerging social effects are monitored and are used to improve the technology and its introduction into society' (67). This kind of piecemeal social engineering (attributed to philosophers Karl Popper and John Dewey) is incremental and focuses on learning from experience, trial-and-error, and beginning with smaller experiments so as to ensure an amount of flexibility and adaptability.

Consider medical experimentation: by understanding and being explicit about the experimental nature of healthcare technologies, an incremental approach is taken to move forward in an ethical way. It would be odd if doctors used a new medication on a group of patients without a methodology for evaluating the performance of the drug, consent of the patients for being a part of the trial, and the use of a control group to isolate variables. Moreover, experiments have mechanisms for monitoring side effects and metrics for success. A proposed blood pressure medication succeeds if blood pressure is lowered and no serious side effects arise. Many of the experiments we perform with AI do not have such mechanisms. Not having these features means that we do not know when the experiment is over, as there is no natural stopping point for which we can re-evaluate the proposed technology's acceptability.

The advantages of conceptualising the release of AI technology into society as an experiment come in many forms. To begin, it raises a whole set of unique ethical issues related to the notion of experimentation in society, on citizens. In a 2007 report by the *European Expert Group on Science and Governance* the authors noted that, 'if citizens are routinely being enrolled without negotiation as experimental subjects, in experiments which are not called by name, then some serious ethical and social issues would have to be addressed' (73).

Next, it demands that an experimental mindset be the guide for the real world use of AI/ML and as such, there are certain restrictions that must be placed on applications of AI. For van de Poel, there are both ethical and epistemological concerns to be addressed;

> *when a new technology is introduced into society it amounts to a de facto social experiment because even if all reasonable efforts to anticipate social consequences haven been undertaken, it is possible, and even likely that there will be unanticipated social consequences. This de facto experimentation can be turned into a mode of more deliberate and responsible experimentation, for example, by*

> *following Popper's idea of piecemeal social experiments. Such responsible experimentation needs to meet both epistemological and ethical constraints. Epistemological constraints are important to ensure learning from social experiments. Ethical constraints are important because these experiments take place in society and may seriously harm individuals as well as society as a whole (67).*

Thus, according to van de Poel (67), not only is it accurate to conceive of new and emerging technologies as social experiments, but it is also prudent to do so. Reliance on an experimental model requires both mindfulness of ethical constraints, and assessment of epistemological constraints (i.e. the risks and benefits resulting from the technology). This becomes particularly essential when AI/ML is applied in contexts that directly affect human lives, such as in government and public institutions dealing with heavy workload and limited resources. Consider for example the use of AI in predicting the maltreatment of children: 'Every year there are more than 3.6 million referrals made to child protection agencies across the US. The practice of screening calls is left to each jurisdiction to follow local practices and policies, potentially leading to large variation in the way in which referrals are treated across the country (2). Perhaps AI/ML tools 'can augment or replace human judgments, which themselves are biased and imperfect' (2). If this were true, AI/ML in such a context could help many children. This is a big 'if' – one that needs to be studied as an experiment with human oversight, that recognises MLs' uncertainties, and that protects the vulnerable demographic involved. As in any experiment, individuals impacted by AI/ML must be protected from harm, and benefits must be explicitly demonstrated by whomever is implementing the AI/ML.

Conceptualising the introduction of technology into society as a social experiment offers an interesting tool for ensuring adaptability of technologists (i.e. those developing the AI/ML algorithms) and technology implementers (i.e. those using the AI/ML solutions in public institutions). Of equal importance, such a precautionary approach also offers protections to citizens impacted by the technology who may have never used it, but who are still impacted by its use in society (e.g. many of the individuals impacted by the catastrophic nuclear disaster in the Chernobyl nuclear power plant in Ukraine were bystanders so to speak, never having used or come into contact with nuclear power or its uses).

In short, when there is much that is unknown about a technology, it should only be allowed under (highly) constrained circumstances that will give us the information necessary for making more concrete policies and legislation around the technology at hand.

Insight #7

- The proliferation of AI in society is an ongoing social experiment full of risks and hypothesised benefits.

Key Insights

Having taken a closer look at what the ethics and technology scholars provide us with, we can see many key lessons that are relevant and necessary for the AI space:

1. Transparency of AI algorithms can mean three distinct things: first, the complexity of modern AI systems – leaving lay users in the dark; second, the intentional obfuscation by those designing AI solutions – leaving lay users and policy-makers in the dark; and third, the inexplicability regarding how a particular input or inputs result in a particular output or outputs – leaving everyone in the dark.
2. Bias and fairness of AI/ML algorithms resulting from the training data is a significant barrier to the ethical development and use of AI/ML. Important questions concerning what is 'fair', what is 'accurate' and how to balance trade-offs between the two must be analysed.

3 AI should be understood as a socio-technical system and should be assessed according to the society in which it has been created. Society's role in the development and applications of AI/ML should not be underestimated.
4 AI can be designed and implemented in ways that obfuscate attributions of responsibility and accountability, but that does not necessarily mean that responsibility and accountability are not possible in the context of AI.
5 Risk assessments, while valuable, do not capture important ethical risks which may be unquantifiable, qualitative, or unobservable.
6 Ethical technology assessments are a viable mechanism for uncovering novel ethical issues which may arise due to the development and use of AI.
7 The proliferation of AI in society is an ongoing social experiment full of risks and hypothesised benefits.

The task now is to understand the implications that these findings have on the current policy landscape of AI, as well as future policy for AI. Before taking this step, let us first address some of the main ethical issues concerning AI raised to date.

5. From AI ethics to AI governance

5.1. The ethics of AI ethics principles

As we have seen, the way in which many institutions and organisations have thought to mitigate some of these ethical issues is to create ethical principles as guidelines for best practice. The principles, then, serve as a voluntary outline for responsible practice. It is important to take note of some of the critical attacks on the development of AI ethics guidelines to date – namely that there are a host of ethical concerns raised about having ethical guidelines and/or principles to guide AI development instead of AI regulation. Some critics believe that ethics-driven self-regulation processes within companies can take the place of external regulation. Others note that the principles do not specify how consumers and citizens will know if and how a company follows AI ethics principles. Still others have raised concerns of 'ethics washing', that ethical discussions have been hijacked by industry, who are acting to prevent the creation of protective governance mechanisms. Each of these criticisms misunderstands the role of ethical governance. Ethics is not intended to replace regulation; rather, ethics is intended to understand the kind of regulation needed. This can only happen, however, when ethics is able to accompany the development of AI. The next part of this study aims at showing the governance mechanisms that ought to be put in place to harness the benefits of AI ethics – to ensure that AI ethics is able to uncover ethical issues in a timely manner as a necessary step towards creating appropriate legislation.

5.2. Where ethics and regulation meet

Ethics is often used to inspire the creation of new regulatory instruments, to call for revisions in existing legislative instruments, or to abolish outdated legislation. As such, ethics plays a motivating role at multiple moments in the regulation and legislative process. This dynamic role for ethics fits nicely with the continuous nature of ethical inquiry and reflection (see sections 4.1 and 4.2) – when technologies provide society with new capabilities, new ethical issues arise that must be dealt with.

Whereas ethics is about searching for broad answers to societal and environmental problems, regulators must make concrete decisions about the future while balancing multiple expert opinions and views. What regulation can do that ethics cannot is codify and enforce ethically desirable behaviour – that we should treat others with dignity and respect, for example, is enshrined in the fundamental rights of the European Union, yet the basis for this comes from centuries of ethics debate and teachings on the inherent moral status and worth of individuals in virtue of their being human. 'Regulation or regulatory governance is, in essence, a form of systematic control intentionally aimed at addressing a collective problem' (15). Regulation in the AI/ML debate must, according to this study, yield a form of systematic control to address the collective problem of the real-world experimentation of AI in society without clear ethical constraints.[23]

In view of framing AI as a social experiment, one should also understand regulation as the attempt to manage the risks associated with the goal of achieving a public set of objectives (74). Thus, regulation for AI in the public domain must be aimed at protecting the foundational elements of liberal democratic societies (the objective) through regulation that insists on ethical constraints and proper experimental methodology to manage (and constrain) risks.

Regulatory instruments have traditionally focused on static technical artefacts and/or the implementation of static regulatory instruments. The speed bump is an often-cited example of a regulatory instrument aimed at controlling the speed of cars on streets where vehicle traffic goes

[23] To be sure, there are other studies conducted by organisations such as the AINOW Institute or ACCESS NOW that call for specific legislative tools to ban certain applications. This study here holds a different aim, notably, to show a broader role for the ethics of technology in the AI Ethics and Regulation debate.

(15, 75). But such an instrument, the speed bump, works in an environment in which we know and understand how the car and the car driver work and, as such, the instrument is based on these variables. The question that AI/ML raises is: What happens when we cannot predict the outcomes of the AI system? How can one regulate said system? Moreover, what happens when the environment and the artefacts in need of regulation are networked, data-driven digital entities rather than analogue artefacts physically embodied in the world? There are no easy answers to these questions, but one responsible way forward is to allow for varying degrees of uncertainty – only if these are documented in a way that is explicit and transparent allowing for a higher degree of accountability on the part of developers and users (i.e. public organisations).

5.3. Regulating the experiment vs regulating the technology

Thinking of technology as a social experiment is not exclusive to the ethics and technology domain; this concept (or a similar version) has been proposed in a variety of sectors, including genetically modified crops.[24] The strengths in shifting the production and use of AI/ML in society from an implicit experiment to an explicit one are many: provide citizens, users and customers with the knowledge that they are involved in an experimental process; provide opportunities to explore less ethically harmful forms of experimentation; provide moments for deliberate ethical learning (as opposed to a strict focus on technical learning), to name a few.[25]

The framing of AI/ML as a real world experiment shifts the question of evaluating the moral acceptability of AI/ML in general to the question of 'under what conditions is it acceptable to experiment with AI/ML in society?' Moreover, this also allows the delineation of certain application domains or stakeholder groups for which/whom it becomes unacceptable to deploy AI/ML; for example, if the risks to children's and/or refugees' welfare outstrip the chances of a benefit, then AI/ML may be deemed unacceptable for that application and/or group. It is only possible, however, to understand the risks to vulnerable demographics, or the risks associated with particular technological capabilities, when all such considerations are made explicit. By framing AI in society as a deliberate experiment, it becomes possible to log, track, and deliberate said risks (both quantitative and qualitative) as a means for encouraging both technical and moral learning as AI is developed and used in society.

The reasons to legislate on ethical constraints in the experiment of AI/ML in society are multiple: the criticality of ethical and human rights issues raised by AI development and deployment; the need to protect people (i.e. the principle of proportionality); the interest of the state (given that AI will be used in state-governed areas such as prisons, taxes, education system etc.); the need to create a level playing field (e.g. self-regulation is not enough to ensure that all organisations making AI/ML will do so in the same way); the need for the development of a common set of rules for all stakeholders to uphold; and the protection from negative outcomes that may result from this new and emerging technology.

5.4. Policy options for governing AI through ethical constraints

Given that there is reason enough to believe that AI can be beneficial, if developed and used properly, it must also be developed in an environment conducive to responsible development. Thus, the goal of this section is to outline a common set of ethical constraints for all public and government organisations purchasing or using AI/ML algorithms, as well as private companies who are developing AI/ML algorithms that will be sold for use in governments and public institutions. In

[24] In the case of genetically modified (GM) crops, agriculture biotechnology in Europe has even been framed in terms of 'technological development as a real-world experiment' (76).

[25] For a more detailed discussion on technology as a social experiment see van de Poel, Asveld, Mehos book 'New Perspectives on Technology in Society' (2017).

so doing, the aim of these ethical constraints as governance mechanisms is to create an environment which encourages the responsible development of AI socio-technical systems. To be sure, **the intention is not to exclude private organisations from these or other policy options**; rather, the goal is to *begin* with public organisations. Accordingly, private companies are no longer considered private when providing a service that falls under the domain of public services (e.g. policing, education etc.). This document is intended to tackle the public sector as a first step, the next step will be to outline the policy options for the private sector specifically.

The ethical constraints centre on the idea of logging essential steps and decision-making in the AI/ML life cycle: 'Good practice in AI comes down to logging – that is, maintaining records about procedures followed in the development and operation of the AI system' (5). The logging discussed in this study has to do with the ethical considerations relevant to the technology, and ultimately, the creation of a systematic procedure for tracking risks as a requirement to engage in ethical evaluations and furthermore, for thoroughly documenting such ethical evaluations. The ethical constraints follow directly from the key insights attributed to the ethics of technology as a field of study. As such, the following policy options chart a path towards accountability insofar as procedures and decisions of an ethical nature are logged and this is done based on certain findings from the ethics of technology.

To remind the reader, the following insights are used to ground the policy options:

1 Transparency of AI algorithms can mean three distinct things: first, the complexity of modern AI systems – leaving lay users in the dark; second, the intentional obfuscation by those designing AI solutions – leaving lay users and policy-makers in the dark; and third, the inexplicability regarding how a particular input or inputs result in a particular output or outputs – leaving everyone in the dark.
2 Bias and fairness of AI/ML algorithms resulting from the training data is a significant barrier to the ethical development and use of AI/ML. Important questions concerning what is 'fair', what is 'accurate' and how to balance trade-offs between the two must be analysed.
3 AI should be understood as a socio-technical system and should be assessed according to the society in which it has been created. Society's role in the development and applications of AI/ML should not be underestimated.
4 AI can be designed and implemented in ways that obfuscate attributions of responsibility and accountability, but that does not necessarily mean that responsibility and accountability are not possible in the context of AI.
5 Risk assessments, while valuable, do not capture important ethical risks which may be unquantifiable, qualitative, or unobservable.
6 Ethical technology assessments are a viable mechanism for uncovering novel ethical issues which may arise due to the development and use of AI.
7 The proliferation of AI in society is an ongoing social experiment full of risks and hypothesised benefits.

5.4.1. Policy option #1

- *It is proposed that all AI/ML system developers be required to have a data hygiene certificate (DHC) to be eligible to sell their solutions to government institutions and public administration bodies.*

Following from Insights #1, #2, #3, #4, and #7 comes the recommendation that **all governmental bodies using AI/ML algorithms and all private companies developing AI/ML algorithms receive, and produce upon request, a data hygiene certificate (DHC)**. Such a certificate acts as a log of the earliest stages in the AI/ML system development (77).

A main concern for implementers purchasing off-the-shelf software is a lack of knowledge of: how the data was acquired (e.g. were users given the option to opt-out of collection, were they informed about how their data would be used?); how the data was sourced (e.g. did a company scrape the internet to acquire the data or was it gathered from the company's own logs?); how the algorithm was trained (using what data); and/or whether any additional tests were carried out of the data (e.g. have the data been explored according to domain-specific aspects to 'highlight the key ingredients in a dataset such as meta-data and populations, as well as unique or anomalous features regarding distributions, missing data, and comparisons to other "ground truth" datasets'?)[26].

It is well known that the quality (or hygiene) of the data plays a key role in the efficacy and accuracy of the algorithm (for more on this see Section 3). Without accurate algorithms, the autonomously developed rules will also be skewed. Consequently, a first ethical constraint is to ensure the quality of the data being used to train the algorithm, where quality is measured according to its sourcing, acquisition, diversity, and labelling. This constraint is also meant to re-calibrate the asymmetry in power between data owners (the creators of websites that collect and track data) and data producers (the citizens and/or consumers using services designed to collect their data).

A DHC is an accreditation of good data provenance models followed by the organisation providing the data and/or developing the AI system and observed (or audited) by a third party. The DHC recommendation is a requirement for organisations to account for: the origin of the data (was it scraped from the internet or collected through a healthcare system etc.); its labelling, storage, sharing, dissemination and anything else relevant to the hygiene of the data. Organisations collecting and labelling data from other sources (e.g. from other companies, institutions, social networking sites etc.) must account for these practices and in doing so be provided with a certificate (by an external regulatory body) once this is complete. Organisations using their own data (e.g. banks, healthcare institutions, schools, governments) must also be able to account for their own collection, storage and use practices through a DHC.

Public organisations even further along in the development chain, i.e. those who are buying off-the-shelf AI software, are required to purchase only those products that can indicate a DHC in the development of the AI software being purchased. This step can be executed in a variety of ways, including, for example, 'The Data Nutrition Project,' which is 'a diagnostic framework that lowers the barrier to standardised data analysis by providing a distilled yet comprehensive overview of dataset "ingredients" before AI model development'.[27]

The goal of this policy option is to change the current practices within the data gathering, sharing, and labelling domains. Effectively, the DHC places the burden of proof on data collectors, labellers, and owners to ensure they have followed best practices. Such a certificate does not require insight into the proprietary aspects of the AI system (i.e. companies do not have to divulge their algorithm) and, of equal importance, such a certificate does not require organisations to share their data sets (which may be their source of income) with competing organisations. It will, however, require that systems are put in place to be able to check how data was being collected, by whom, when, was it able to be corrected, redacted, etc. Ultimately, the goal of this policy option is to mandate responsible practices for the very first step in the lifecycle of the AI system: the data.

5.4.2. Policy option #2

- *It is proposed that all public and government organisations using AI systems be required to conduct an ethical technology assessment (eTA) prior to deployment of the AI system.*

[26] See the Data Nutrition Label Project at https://datanutrition.media.mit.edu/

[27] For more on this see https://datanutrition.media.mit.edu/ (Retrieved July 9, 2019)

Following from Insights #1, #2, #3, #4, #5, #6, and #7 comes the recommendation that each **public administration and governmental organisation that will deploy an AI/ML solution be required to conduct an ethical technology assessment (eTA) before deployment.**

Whereas the original vision of an eTA was meant as continual discussion between technologists and ethicists (see Section 4.2.4) the use of the eTA for this recommendation is to mark the second step in the documentation of the AI/ML social experiment (the first being the DHC). Of course, continual dialogue between technologists, ethicists, and public administration officials is desirable; however, this policy option is to ensure a basic minimum ethical constraint – that consideration of ethical risks has been carried out prior to the implementation of the AI/ML and further that knowledge of such risks has been documented.

The eTA is a written document intended to capture and log the dialogue that occurred between ethicist and technologist and/or ethicist and officials of the public administration about to implement the AI/ML solution. The eTA is a list of ethical issues related to the AI/ML application, made by an expert trained to engage in ethical reflection (or at the very least one who is able to envision possible moral risks related to the implementation of the AI/ML). The eTA is the moment where one must consider all the possible ethical risks that could result from the AI/ML application in question. It is also the place where application-specific or demographic-specific issues will be raised – for example, extra precautions if children, elderly people, migrants, or other vulnerable groups are involved should be flagged here. When vulnerable groups are involved in other kinds of human (medical) experimentation, additional measures must be followed, such as having proxy consent if individuals are underage.

In contrast with fundamental research of AI systems for which no application may be intended, when AI systems are deployed in public administration organisations, in governments or in the private sector, there is an intended goal, e.g. to make predictions regarding performance of employees, on citizens' risk of breaking the law, or on parents' ability to parent. In these instances, it should be required to engage in an eTA regarding the ethical impacts that AI systems might have on the practice in which it is being deployed. Thus, the eTA must also include descriptive information about the particular application: the context in which the AI will function (e.g. law, policing, education, healthcare, etc.); the direct and indirectly impacted stakeholders (e.g. in a healthcare context this may be physicians, patients, family of patients, hospital support staff, etc); the practice for which the AI will be used (e.g. disease detection); the type of AI methodology (e.g. good old-fashioned AI, deep learning, neural networks, convolutional neural networks, etc.); and an account of ethical values and human rights in need of attention (e.g. right to a private life, autonomy and ethical values such as privacy, integrity, dignity etc.). These details allow for the possibility to target ethical issues on a fine-grained level and work to avoid discussions of 'random' or implausible issues (21, 78, 79).

An example of an eTA log can be taken from the canonical paper of Palm and Hansson (65), in which they present a case in favour of eTAs, as well as a list of criteria taken from historical evidence of common problem areas.[28] The list of ethical issues to begin the eTA are: 1. dissemination and use of information; 2. control, influence and power; 3. impact on social contact patterns; 4. privacy; 5. sustainability; 6. human reproduction; 7. gender, minorities and justice; 8. International relations; and 9. impact on human values (see section 4.2.4 for more on this). The eTA document is thus a detailed account of how the application may impact all of these nine criteria. Alternatively, an organisation may opt in favour of structuring the eTA according to a set of ethical principles previously developed for AI/ML in particular. Take for example the Ethics Guidelines for Trustworthy

[28] This is also a moment in which a set of AI ethics principles may be used as the framework for conducting the eTA, for example the Guidelines of Trustworthy AI may be used to evaluate a new AI/ML application and logged as the eTA.

AI developed by the High-Level Expert Group on AI.[29] In this case, an organisation could use the seven key requirements to structure an eTA.[30]

The eTA log is meant to put on record that governments and public administration institutions were informed of the possible ethical risks associated with the AI/ML system they are about to, or have, implemented, including the qualitative and the undetectable elements concerning 'risk'. This report does not require them to act on the issues; it is instead meant to raise the issues and to provide an account of the ethical issues that have been documented.

Of equal importance to having the logs in the first place is that the logs must be trustworthy (5). For this to happen there is a need for measures both technical (e.g. cybersecurity) and practical (e.g. use of objective, external third parties as assessors).

Not only will the continuous process of eTAs for government applications be beneficial to ensure transparency and accountability, but they will also allow for the tracking of values over time, i.e. changes in value prioritisation, interpretation, and assessment. It may also be important to note how certain risks are more prevalent for one application over another, or how the interpretation and prioritisation of risks shifts over time as we gain a deeper understanding of AI/ML. As such, the report is a documentation of the ethical concerns accompanying the real-world application of AI/ML in society but also society to understand and study AI as a socio-technical system.

5.4.3. Policy option #3

- *It is proposed that all public administration institutions and government bodies be required to show clear goals of the AI/ML application.*

Following from Insights #1, #3, #4 and #7, it is unethical to experiment on individuals and/or society without a clear idea of the kinds of impacts such experimentation might have, and without a goal to measure the efficacy of the intervention (in this case the AI/ML algorithm). In order to address such gaps, it is proposed that **AI implementers should not be allowed to deploy AI/ML applications without a clear idea of what they are trying to achieve and what metrics they are using to measure success or failure**. By this policy option, it is not possible to deploy AI/ML in society in the hope of learning an unknown 'something'. Rather, there must be a need for a specific and explicit 'something' to be learned. The specific aim and scope of the AI/ML experiment must also be stated as part of the eTA (see above).

Note, this policy option does not mean that researchers or industry leaders will not also learn new things; rather, any exploratory work must be done in the context of a defined target, and this policy option is meant as a way to measure the efficacy of reaching said target.

Of equal importance, this 'something' to be learned should be clearly aligned with the stated benefits to the subjects involved. It should not be possible to deploy AI/ML algorithms in society without stated benefits for the subjects involved (note: 'subjects' is not the same as the persons executing the experiment but the individuals who will be participants in the experiment). Moreover, benefits from one application should not be used to justify the risks in another application (see section 4.2.3). If no benefits can be identified for a specific AI/ML application then serious reconsideration of the AI/ML deployment must be undertaken.

[29] For more on this, see https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[30] It should be noted that the current assessment list focuses on the practical implementation of AI/ML in an organisation rather than on the ethical consequences of said implementation. As such, qquestions would need to be created to translate the key requirements towards the possibility for ethical risks, e.g. in a healthcare context in which AI/ML systems are used to assist in decision making, a question concerning human oversight may include questions such as: 'what is the long term impact on human oversight if decisions by medical practitioners are encouraged to be accompanied by an ML system for which explainability of the output is not available?'

5.4.4. Policy option #4

- *It is proposed that organisations deploying AI systems should produce an 'accountability report' in response to the eTA.*

Following from Insights #1, #2, #3, #4, #5, #6, and #7 comes the recommendation that **organisations deploying AI/ML must also provide an 'accountability report' in which they address any flagged ethical concerns regarding the AI/ML deployment raised in the eTA.**

The accountability report is the third step in logging the AI/ML usage in public administration, in government and in the private sector. Whereas the eTA is meant to draw out the possible negative consequences of implementing an AI system (completed by an external third party), the accountability report is a response to this report, completed by the organisation implementing the AI/ML system. It is meant as a response to the ethical and human rights issues that were identified in the eTA. Thus, in the accountability report, institutions will be required to account for how they have mitigated or corrected the concerns raised in the eTA.

This is also a moment in which the various ethical theories (aka conceptual frameworks) are meant to assist in the decision-making process; for better or for worse, organisations can indicate which conceptual framework (or combination of conceptual frameworks) was used as part of the decision-making process.

Engaging in the practice of eTAs, coupled with accountability reports, also acts as a reinforcement of the continuous practice of ethics – these reports are not a checklist but are a series of steps to raise ethical issues, acknowledge them, and take steps (on the record) to show an attempt to minimise them.

Given that it is not possible to legally require organisations to follow ethical advice, it is not possible to mandate that all ethical concerns raised in the eTA are taken into account. However, if all ethical concerns and risks, on both an individual and collective level, are made public then these records will serve as a means to account as to whether or not the company was informed of the possible risks, and what their response to these risks was at the time.[31] In this way, it no longer becomes possible for a company to claim negligence about the possible risks of a certain AI/ML application.

[31] Such a process follows the same basic approach as a risk assessment in auditing, for example; risks are brought to the client after an internal review of the system, and it is up to the client to decide if and how they will respond to these.

6. Anticipating problems

As with any set of policy options for policy-makers, there are sure to be concerns about the scope of the policy options and/or the ability to implement said policy options. So, too, will concerns be raised about the policy options proposed here. This section is an attempt to anticipate a number of possible problems with the above policy options and to provide a response to said problem.

6.1. Who is going to make eTA reports? Or, the role of the ethicist in our future with AI

Some readers may wonder: who should be tasked with completing the eTA? Can this be delegated to any member of the design team, or are there certain qualifications that the individual tasked with completing the eTA should have?

In the same way that we cannot and should not expect the expert graphic designer to be an expert in corporate negotiations, we should not expect the data scientist to be an expert in ethics. As Palm and Hansson (65) accurately write, 'Engineers are seldom trained to discuss the ethical issues in a pre-emptive perspective. The training needed would be that of identifying consequences for different stakeholders at an early stage, that is, to identify the potential problems' (p. 547). The tasks in the eTA log require experience in identifying ethical issues and placing them within a conceptual framework for analysis. These are the skills of (applied) ethicists; a data scientist or engineer is unlikely to possess them. Thus, we must acknowledge a presumed future role for ethics and for ethicists in the regulation process engaged in organisations around AI/ML.

6.2. Will SMEs be able to afford this?

In the case of small and medium enterprises (SMEs), with three to four people working in the organisation, will it be a problem to insist that an eTA and an accountability report be completed? And further, if they need to hire an external third party to do it, how will they afford this? In order to create a level playing field across SMEs, there should be an opportunity within the first three years after such a policy option is made into policy that offers small and medium companies EU funding to assist them with report completion as well as with the necessary capacity-building measures. This model parallels the incremental process for companies to comply with the General Data Protection Regulation (GDPR) – the data privacy regulation harmonising data privacy laws across Europe.[32]

6.3. Should all governments and/or public administration organisations be subject to regulation?

Given the labour involved in organisations completing eTAs and accountability reports, may certain organisations with 'benign' applications of ML/AI be exempt from a regulatory process? The response to such a critique would be to ask: what is the harm in engaging in an eTA for an AI product, even if the AI/ML application appears to be harmless? And does that harm (i.e. the time spent) have more weight than the good of assuring that a company follows ethical industry practices? Is it not better to have a publicly available report for any company, rather than the alternative, i.e. a pre-emptive 'belief' in no harm only to learn later there was unanticipated harm?

[32] For more, see https://eugdpr.org/

6.4. AI is already out there

There is, of course, the concern that because AI/ML algorithms are already pervasive in society, it is too great a challenge to attempt to regulate them after that fact. However, one need only look to the creation of the GDPR to ask if it is an impossible task to regulate an already pervasive technology and/or practice in society. The GDPR replaced the Data Protection Directive 95/46/EC with the intention to 'reshape the way organisations across the region approach data privacy'.[33] The GDPR initiative shows that not only is it possible to regulate a product after the fact, it is essential for the protection of citizens in view of the new risks and harms that arise through digital technologies. Moreover, the spirit of the GDPR is also to contrast the tendency towards technological determinism (see section 5.2.2) – technology and/or the tech industry are not allowed to dictate the terms of how technology ought to impact citizens, human rights, and society at large. **The fact that technology often develops faster than ethical guidelines can be made does not excuse industry or public institutions from regulation**. Furthermore, there are ways in which companies can work to achieve these policy options gradually, e.g. by setting benchmarks and allowing reasonable timelines to achieve them; complete transition does not have to happen overnight.

6.5. Ethics stifling innovation

There is a familiar push-back on the ethics of technology that proposes that ethics stifles, or even prohibits, innovation. This 'slowing down' of technology development need not be the case if ethicists are working within the organisations that are using AI, or are up-to-date with the latest in AI/ML techniques and applications. Ethicists with insufficient context may provide guidelines that are impractical. The solution is not to disregard ethics; rather, the solution is to invite ethicists into the development process to facilitate applicable feedback. In a collaborative environment, conversations between developers and ethicists can even act to inspire new kinds of innovations – AI/ML solutions that are created with the wellbeing of users in mind from the very beginning. Consider the sustainability movement as a parallel; the concept of sustainability, while once considered a constraint that stifled innovation, is now a constraint used to push new kinds of innovation along the paradigm of sustainability (e.g. circular economy, minimising carbon footprint, recycling, minimising waste, etc.). Because there is pressure on organisations to work within these constraints (insofar as consumers are willing to purchase products from such purpose-driven sustainable companies), they find new ways to move forward. AI/ML developers, working under ethical constraints such as those proposed in this study, have the opportunity to seek new methods and applications to meet these constraints.

[33] See https://eugdpr.org/

7. Conclusion

The last three years have seen an unprecedented number of AI/ML applications, along with an equally unprecedented number of AI ethics principles developed by governments, civil society, private companies, and multi-stakeholder groups. Although there has been an increase in attention paid to ethics in the AI debate, little is done to show what ethics means for the creation of policy and regulation of AI. *What ethics provides as a field of study is a vision of the future: a normative perspective, rather than descriptive, with an eye to the 'good life'*. The question at the centre of this study was to ask: **How can we move from AI ethics to specific policy and legislation for governing AI?**

With this focus, the study addresses the features of AI/ML as a unique technology allowing for AI ethics to also have a unique focus. This study illustrates the role that ethics as a discipline can have in understanding the reciprocal relationship between AI as a technology, as well as the social systems within which AI is created, and into which it is disseminated (i.e. AI as a socio-technical system). Of equal importance, this study shows how the appropriate way to frame the real-world use of AI today is to understand it as a social experiment. Given the lack of operational experience we have with AI – the level of uncertainty and risk – it is wise to introduce experimental conditions for the real-world applications of AI, especially when it comes to ethical constraints and requirements for demonstrating clear benefit.

A cross-cutting theme in all of this was acknowledging the temptation towards technological determinism – that technology's capabilities determine if/how we regulate it – but remembering that AI as a socio-technical system (along with the embedded values approach) means recognising the role society plays in shaping the technology and the regulation thereof. Science and technology studies teach us that 'technology development involves many different actors with interests that push development in a variety of directions … the many actors affect the direction of development' (4). Given this plurality of involvement, it is equally important to remember that 'there is a lot more than technical feasibility involved in shaping future technologies and, most importantly, the outcomes of research and development are contingent, not inevitable' (4). On a related note, what is also often missing from the discussion is that future technologies are not the result of technical feasibility alone, but instead require human activity in order to become feasible. Researchers must study the variety of ways in which AI can be developed, funding instruments must be put in place to support research and development, developers must choose between one design paradigm and another. In other words, the feasibility of AI is not something to be discussed in a vacuum, and AI will happen only at the hands of humans working to make it happen.

Together, these insights point towards specific policy options to regulate the experimental use of AI in society. These policy options are the ethical constraints under which AI/ML applications can be deemed acceptable upon their introduction into society. The policy options presented here require a series of logging records of important steps so as to ensure accountability and transparency through the implementation of AI/ML solutions in governments and public organisations. Moreover, this logging is the first step in allowing ethics to play a significant role in the implementation of AI for the public good. The only question remaining then is: Who will make this happen and will it be in time?

REFERENCES

1. Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A. Algorithmic decision making and the cost of fairness. ArXiv170108230 Cs Stat [Internet]. 2017 Jan 27 [cited 2019 Jul 9]; Available from: http://arxiv.org/abs/1701.08230
2. Chouldechova A, Putnam-Hornstein E, Benavides-Prado D, Fialko O, Vaithianathan R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. :15.
3. Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S. Runaway Feedback Loops in Predictive Policing. ArXiv170609847 Cs Stat [Internet]. 2017 Jun 29 [cited 2019 Jul 9]; Available from: http://arxiv.org/abs/1706.09847
4. Johnson DG. Technology with No Human Responsibility? J Bus Ethics. 2015 Apr; 127(4):707–15.
5. Bryson J. A smart bureaucrat's guide to AI regulation [Internet]. 2019. Available from: https://joanna-bryson.blogspot.com/2019/01/a-smart-bureaucrats-guide-to-ai.html
6. Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. ArXiv160905807 Cs Stat [Internet]. 2016 Sep 19 [cited 2019 Jul 9]; Available from: http://arxiv.org/abs/1609.05807
7. Tene O, Polonetsky J. Taming The Golem: Challenges of Ethical Algorithmic Decision-Making. N C J Law Technol. 2018;19(1):50.
8. Andersen L. Human Rights in the Age of Artificial Intelligence. [Internet]. Access Now; 2018. Available from: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
9. Johnson DG, Miller KW. Un-making artificial moral agents. Ethics Inf Technol. 2008 Sep 1;10 (2–3):123–33.
10. Maslen H. Computer vision and emotional privacy | Practical Ethics [Internet]. University of Oxford: Ethics in the News. 2014 [cited 2019 Feb 28]. Available from: http://blog.practicalethics.ox.ac.uk/2014/03/computer-vision-and-emotional-privacy/
11. Pasquale F. The Black Box Society: The Secret Algorithms That Control Money and Information [Internet]. Cambridge, Mass.: Harvard University Press; 2015 [cited 2017 Jun 2]. Available from: http://www.jstor.org/stable/j.ctt13x0hch
12. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C, et al. Machine behaviour. Nature. 2019 Apr;568(7753):477.
13. Santoni de Sio F, van den Hoven J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. Front Robot AI [Internet]. 2018 [cited 2019 Jul 7];5. Available from: https://www.frontiersin.org/articles/10.3389/frobt.2018.00015
14. Whittaker M, Crawford K, Dobbe R. AI Now Report 2018 [Internet]. AI Now; 2018. Available from: https://ainowinstitute.org/AI_Now_2018_Report.pdf
15. Yeung K. 'Hypernudge': Big Data as a mode of regulation by design. Inf Commun Soc. 2017 Jan 2;20(1):118–36.
16. Beauchamp TL, Childress JF. Principles of biomedical ethics. Oxford; New York: Oxford University Press; 2001.
17. Adamson G, Havens JC, Chatila R. Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems. Proc IEEE. 2019 Mar;107(3):518–25.
18. Dignum V. RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES. Daffodil Int Univ Repos. 2017 Sep;(1):9.
19. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. Big Data Soc. 2016 Dec 1;3(2):2053951716679679.
20. Friedman B. Value-sensitive Design. interactions. 1996 Dec;3(6):16–23.
21. van Wynsberghe A. A method for integrating ethics into the design of robots. Ind Robot Int J. 2013 Aug 16;40(5):433–40.

22. Alshammari M, Simpson A. Towards a Principled Approach for Engineering Privacy by Design. In: Schweighofer E, Leitold H, Mitrakas A, Rannenberg K, editors. Privacy Technologies and Policy. Springer International Publishing; 2017. p. 161–77. (Lecture Notes in Computer Science).
23. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. [Internet]. IEEE; 2018. Available from: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
24. van de Poel I. Design for value change. Ethics Inf Technol [Internet]. 2018 Jun 26 [cited 2019 Jul 13]; Available from: https://doi.org/10.1007/s10676-018-9461-9
25. Reisman D, Schultz J, Crawford K, Whittaker M. ALGORITHMIC IMPACT ASSESSMENTS: [Internet]. 2018 [cited 2019 Jun 11] p. 22. Available from: https://ainowinstitute.org/aiareport2018.pdf
26. Russell SJ, Norvig Peter. Artificial intelligence : a modern approach. Englewood Cliffs, N.J.: Prentice Hall; 1995.
27. Turing AM. Computing machinery and intelligence. Mind Q Rev Psychol Philos. 1950;LIX(236):433–.
28. Agrawal A, Gans J, Goldfarb A. Prediction Machines: The Simple Economics of Artificial Intelligence. Harvard Business Press; 2018. 357 p.
29. High-Level Expert Group on AI. A definition of Artificial Intelligence: main capabilities and scientific disciplines [Internet]. Brussels: European Commission; 2018 Dec [cited 2019 Feb 20]. (Policies, Information and Services). Available from: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
30. Hill RK. What an Algorithm Is. Philos Technol. 2016 Mar 1;29(1):35–59.
31. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc. 2016 Jun 1;3(1):2053951715622512.
32. Danaher J. Toward an Ethics of AI Assistants: an Initial Framework. Philos Technol. 2018 Dec 1;31(4):629–53.
33. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach. 2018 Dec 1;28(4):689–707.
34. Sharkey N. Cassandra or False Prophet of Doom: AI Robots and War. IEEE Intell Syst. 2008 Jul;23(4):14–7.
35. Taddeo M, Floridi L. How AI can be a force for good: Science. 2018 Aug 24;361(6404):751–2.
36. Robbins S. AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI Soc [Internet]. 2019 Apr 10 [cited 2019 Jul 14]; Available from: https://doi.org/10.1007/s00146-019-00891-1
37. Hwang T. Computational Power and the Social Impact of Artificial Intelligence [Internet]. Rochester, NY: Social Science Research Network; 2018 Mar [cited 2019 Jul 14]. Report No.: ID 3147971. Available from: https://papers.ssrn.com/abstract=3147971
38. Brey P. Ethical aspects of facial recognition systems in public places. J Inf Commun Ethics Soc. 2004 May 31;2(2):97–109.
39. Nemitz Paul. Constitutional democracy and technology in the age of artificial intelligence. Philos Trans R Soc Math Phys Eng Sci. 2018 Nov 28;376(2133):20180089.
40. Holm EA. In defense of the black box. Science. 2019 Apr 5;364(6435):26–7.
41. Haussler D. Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. Artif Intell. 1988;36(2):177–221.
42. Heckmann JJ. Selection bias as a specification error. Econom J Econom Soc. 1979;153–61.
43. Van den Hoven J, Lokhorst G-J, Van de Poel I. Engineering and the Problem of Moral Overload. Sci Eng Ethics. 2012 Mar;18(1):143–55.
44. Mill JS. Utilitarianism. Longmans, Green and Company.; 1895.

45. Poel I van de. Why New Technologies Should be Conceived as Social Experiments. Ethics Policy Environ. 2013 Oct 1;16(3):352–5.
46. Wynsberghe A van, Robbins S. Critiquing the Reasons for Making Artificial Moral Agents. Sci Eng Ethics. 2018 Feb 19;1–17.
47. Vallor S. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press; 2016. 329 p.
48. Winner L. Do Artifacts Have Politics? Daedalus. 1980;109(1):121–36.
49. Akrich M. The de-scription of technical objects. In: Bijker WE, Law J, editors. Shaping technologybuilding society. MIT Press; 1992. p. 205–24.
50. Brey P. Values in Technology and Disclosive Computer Ethics. In: Floridi L, editor. The Cambridge Handbook of Information and Computer Ethics. Cambridge University Press; 2010. p. 41–58.
51. Nissenbaum H. How computer systems embody values. Computer. 2001;34(3):120–119.
52. van de Poel I. Conflicting Values in Design for Values. In: van den Hoven J, Vermaas PE, van de Poel I, editors. Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains [Internet]. Dordrecht: Springer Netherlands; 2015 [cited 2019 Jun 13]. p. 89–116. Available from: https://doi.org/10.1007/978-94-007-6970-0_5
53. van Wynsberghe A. Designing Robots for Care: Care Centered Value-Sensitive Design. Sci Eng Ethics. 2012 Jan 3;19(2):407–33.
54. van Wynsberghe A, Robbins S. Ethicist as designer: a pragmatic approach to ethics in the lab. Sci Eng Ethics. 2014 Dec;20(4):947–61.
55. A van W, C G. Telepsychiatry and the meaning of in-person contact: a preliminary ethical appraisal. Med Health Care Philos. 2009;12(4):469–76.
56. van de Poel I. Values in engineering design. In: Meijers A, editor. Handbook of the philosophy of science Volume 9: Philosophy of technology and engineering sciences. Oxford: Elsevier; 2009.
57. Verbeek P. Morality in Design; design ethics and the morality of technological artifacts. In: Vermaas PE, editor. Philosophy and design : from engineering to architecture. Dordrecht: Springer; 2008. p. 91–102.
58. Bijker WE, Law J. Shaping technology/building society : studies in sociotechnical change. Cambridge, Mass.: MIT Press; 1992. (Inside technology).
59. Ropohl G. Philosophy of Socio-Technical Systems. Techné Res Philos Technol. 1999;4(3):186–94.
60. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol. 2004;6(3):175–83.
61. Santoro M, Marino D, Tamburrini G. Learning robots interacting with humans: from epistemic risk to responsibility. AI Soc. 2008;22(3):301–314.
62. Nagenborg M, Capurro R, Weber J, Pingel C. Ethical regulations on robotics in Europe. AI Soc. 2008;22(3):349–66.
63. van de Poel I, Nihlén Fahlquist J, Doorn N, Zwart S, Royakkers L. The Problem of Many Hands: Climate Change as an Example. Sci Eng Ethics. 2012 Mar 1;18(1):49–67.
64. Hansson SO. Evaluating the Uncertainties. In: Hansson SO, Hirsch Hadorn G, editors. The Argumentative Turn in Policy Analysis: Reasoning about Uncertainty [Internet]. Cham: Springer International Publishing; 2016 [cited 2019 Jun 12]. p. 79–104. (Logic, Argumentation & Reasoning). Available from: https://doi.org/10.1007/978-3-319-30549-3_4
65. Palm E, Hansson SO. The case for ethical technology assessment (eTA). Technol Forecast Soc Change. 2006;73(5):543–58.
66. Allison W. We Should Stop Running Away from Radiation. Philos Technol. 2011 Jun 1;24(2):193–5.
67. van de Poel I. An Ethical Framework for Evaluating Experimental Technology. Sci Eng Ethics. 2016 Jun 1;22(3):667–86.

68. Have HAMJ ten. Medical Technology Assessment and Ethics. Hastings Cent Rep. 1995;25(5):13–9.
69. McCarthy OJ. AI & Global Governance: Turning the Tide on Crime with Predictive Policing [Internet]. United Nations University Center for Policy and Research; 2019 Feb. Available from: https://cpr.unu.edu/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html
70. Understanding Community Policing: A Framework for Action [Internet]. Bureau of Justice Assistance, Response Center; 1994. (Bureau of Justice Assistance). Available from: https://www.ncjrs.gov/pdffiles/commp.pdf
71. Morozov E. To Save Everything, Click Here: The Folly of Technological Solutionism. Reprint edition. New York, NY: PublicAffairs; 2014. 432 p.
72. Zimmer M. 'But the data is already public': on the ethics of research in Facebook. Ethics Inf Technol. 2010 Dec 1;12(4):313–25.
73. Felt U, Wynne B, Callon M, Goncalves M, Jasanoff S, Jepsen M, et al. Taking European knowledge society seriously. 2007.
74. Black J. Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a 'Post-Regulatory' World. Curr Leg Probl. 2001 Jan 1;54(1):103–46.
75. Latour B. Where Are the Missing Masses? The Sociology of a Few Mundate Artifacts. In: Bijker W, Law J, editors. Shaping Technology // Building Society: Studies in Sociotechnical Change. MIT Press; 1992. p. 225–58.
76. Levidow L, Carr S. GM crops on trial: Technological development as a real-world experiment. Futures. 2007 May 1;39(4):408–31.
77. Bryson J. My comments/critiques on the EU's High Level Expert Group on AI's 'ethical guidelines' [Internet]. 2019. Available from: https://joanna-bryson.blogspot.com/2019/02/my-commentscritiques-on-eus-high-level.html
78. Sio FS de, Wynsberghe A van. When Should We Use Care Robots? The Nature-of-Activities Approach. Sci Eng Ethics. 2015 Nov 7;1–16.
79. van Wynsberghe A. Healthcare Robots: Ethics, Design and Implementation. Ashgate Publishing, Ltd.; 2015. 167 p.

There is little doubt that artificial intelligence (AI) and machine learning (ML) will revolutionise public services. However, the power for positive change that AI provides simultaneously has a potential for negative impacts on society.

AI ethics work to uncover the variety of ethical issues resulting from the design, development, and deployment of AI. The question at the centre of all current work in AI ethics is: 'How can we move from AI ethics to specific policy and legislation for governing AI?'

Based on a framing of 'AI as a social experiment', this study arrives at policy options for public administrations and governmental organisations who are looking to deploy AI/ML solutions, as well as the private companies who are creating AI/ML solutions for use in the public arena. The reasons for targeting this application sector concern: the need for a high standard of transparency, respect for democratic values, and legitimacy. The policy options presented here chart a path towards accountability; procedures and decisions of an ethical nature are systematically logged prior to the deployment of an AI system. This logging is the first step in allowing ethics to play a crucial role in the implementation of AI for the public good.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

ISBN 978-92-846-5855-8 | doi: 10.2861/247 | QA-03-19-800-EN-N