DELPHI – INTERDISCIPLINARY REVIEW OF EMERGING TECHNOLOGIES

2019 | VOLUME 2 | NUMBER 4

Delphi

AI Governance

Towards an Index of Digital Responsibility

Eva Thelisson, Jean-Henry Morin and Johan Rochel

AI and Online Intermediation Platforms

Conciliating Economic Efficiency and Ethical Issues

Frédéric Marty and Thierry Warin

AI Safety

Classification Schemas for Artificial Intelligence Failures

Peter J. Scott and Roman V. Yampolskiy

The Legal Publisher
lexxion

DATA PROTECTION INSIDER

BIWEEKLY NEWS ON EU DATA PROTECTION

Sign up lexxion.eu/en/newsletter/

Powered by EDPL and edited by
Dara Hallinan and **Diana Dimitrova**,
this **free newsletter** provides you with
up-to-date information and insights on
Data Privacy, Data Protection and GDPR related issues.

EdpL
EUROPEAN DATA
PROTECTION
LAW REVIEW

www.lexxion.eu

Contents

Publisher
Lexxion Verlagsgesellschaft mbH
Güntzelstraße 63 · 10717 Berlin · Germany
Phone: +49 30/81 45 06-0 · Fax: +49 30/81 45 06-22
www.lexxion.eu

Typeset
Automatic typesetting by metiTEC-software
me-ti GmbH, Berlin

Delphi annual subscription rates 2019 (4 issues)*

Regular Price	
printed version only	417,00 €
printed version + online edition (incl. archive)**	471,00 €
online edition only (incl. archive)**	410,00 €
Startups	
printed version only	165,00 €
printed version + online edition (incl. archive)**	198,00 €
online edition only (incl. archive)**	158,00 €
Students	
printed version + online edition (incl. archive)***	98,00 €
online edition only (incl. archive)***	72,00 €

* Prices include Postage and Handling. EU Member States: VAT will be added if applicable.
** Multi-user online acess via IP authentication (IP-range).
*** Single user online access via user name and password.

Delphi is supplied under our terms of sale and supply.
Copies of our terms and conditions are available upon request.
Lexxion Verlagsgesellschaft mbH. VAT Reg.No. DE 209981990.

For further information please contact
info@lexxion.eu
Tel.: +49 30/81 45 06-0 · Fax: +49 30/81 45 06-22

Contributions
are welcome and should be submitted according to the *Delphi* author guidelines. Any such contribution is accepted on the understanding that the author is responsible for the opinions expressed in it. More informations available at delphi.lexxion.eu.

Photocopying
All rights reserved. No part of this journal may be reproduced in any form whatsoever, e.g. by photo print, microfilm, or any other means, without prior permission of the publisher. This journal was carefully produced in all its parts. Nevertheless, authors, editors and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Ownership and shareholdings pursuant to Section 7 lit. a No. 2 and Section 2 No. 6 of the Berlin Press Act: Shareholder of Lexxion Verlagsgesellschaft mbH is Dr. Wolfgang Andreae, Publisher, Berlin.

This journal may be cited as [2019] Delphi.

ISSN (Print) 2626-3734 · ISSN (Online) 2626-3742

Delphi

EDITOR–IN–CHIEF

Ciano Aydin
University of Twente, the Netherlands

ASSOCIATE EDITORS

Francesca Bosco
World Economic Forum, Switzerland

Sarah Fischer
Deutsche Gesellschaft für Internationale
Zusammenarbeit GmbH, Germany

Florian Krausbeck
ambrite AG, Switzerland

Anna Laesser
Impact Hub Berlin, Germany

Matthias Lamping
Max Planck Institute for Innovation and
Competition, Germany

Vince Madai
Charité Berlin, Germany

Ida Rust
University of Twente, the Netherlands

Yueh-Hsuan Weng
Tohoku University, Japan

Cees Zweistra
Delft University of Technology, the Netherlands

EDITORIAL BOARD

Steffen Augsberg
Justus-Liebig-Universität Gießen, German Ethics
Council, Germany

Woodrow Barfield
Professor Emeritus, Chapel Hill, USA

Aubrey de Grey
SENS Research Foundation, USA

William Echikson
Centre for European Policy Studies, Belgium

Paul Nemitz
European Commission, Belgium

Vuyiswa M'Cwabeni
SAP, Germany

Nishant Shah
ArtEZ University of the Arts, the Netherlands

Stefan Lorenz Sorgner
John Cabot University, Italy

Rob van den Hoven van Genderen
VU University Amsterdam, the Netherlands

Steven Umbrello
Institute for Ethics and Emerging Technologies,
Italy

Anna Zeiter
eBay Inc., Switzerland

EXECUTIVE EDITORS

Clara Hausin
Lexxion Publisher, Germany
hausin@lexxion.eu

Jakob McKernan
Lexxion Publisher, Germany
mckernan@lexxion.eu

Editorial: The Central Role of States for Building a Balanced AI Governance

The disruptive nature of artificial intelligence transforms almost all human activities and requires a cohesive and sustainable AI governance framework on a global scale. This framework should aim at managing both the opportunities and the risks derived from this technology in a proportionate manner. The digital economy has increased the need for a trusted ecosystem, including reinforced regulations and additional constraints for all actors dealing with artificial intelligence at whatever part of the value chain. As a result, public actors have initiated a process that promotes a balanced approach beneficial for all innovation, society and individuals. It is part of a concerted international framework at EU, OECD and G20 level and also includes isolated projects like the 'Model AI Governance framework' from Singapore.

The goodwill of States is key in ensuring an effective governance of artificial intelligence. The peer review mechanism or reviews by independent experts can play a central role in the effective implementation of these frameworks. Depending on how AI will be used, it can indeed either contribute to achieving the UN's Sustainable Development Goals (SDGs) or lead to negative societal externalities like harm to citizens, misuse of data, the manipulation of people (deep fake misuse) or mass surveillance.

Within the framework of their sovereignty of positive responsibility and protection, States are responsible for the implementation of these non-binding principles or guidelines on artificial intelligence at a national level. AI governance safeguards users' interest of digital services and products, as well as citizens' interests in public spaces. One of the most recent examples is the EU project to ban facial recognition technologies for up to 3 to 5 years, following the Clearview scandal. This ban is founded on the General Data Protection Regulation and the right 'not to be subject of a decision based solely on automated processing, including profiling, which produces legal effects'.

While the constraint on States to implement the AI principles and guidelines as coordinated at the international level has not changed in nature, this pressure – mainly political – seems to have increased, as is the case in non-cooperative territories in tax matters.

AI Governance Should Consider the Long Term Perspectives

In the wrestling match between excess of individual freedom and the 'common good', the question as to what constitutes 'meaningful governance' is a pertinent one. The AI Transparency Institute holds the opinion that meaningful governance via a binding, directly enforceable regulation is necessary to ensure the safety of AI. It should be proportionate, based on a risk-based approach and respect democratic values and prin-

DOI: 10.21552/delphi/2019/4/3

ciples. Integrating long term criteria within decision making processes would also contribute to mitigating risk and ensuring irreversible steps are avoided.

In particular, ex-ante and ex-post mechanisms should be built for a sustainable AI governance, mainly based on accountability, transparency, good design, safety and liability. Regular risk assessments for high risk projects prior to market deployment of digital products and services and legally binding instruments should be put in place to safeguard the democratic use of artificial intelligence. Soft law like quality labels and certification mechanisms could complete this framework.

In a global market economy, not bound by territorial borders and mainly driven by interdependencies and short term indicators like stock exchange value or polling data, long term interest of future generations should be a key component of the AI governance framework.

Call for Multilateral Governance, Including all Private and Public Stakeholders

Our focus in this Special Issue is on the governance of AI and on Governance via AI. We recommend a hybrid governance methodology by State (hard law) and by the market (soft law), inspired by the GDPR and the 108+ Convention, in particular a network of independent control authorities and effective legal remedies (class actions).

In the first contribution, *Mael Pegny* highlights the need to recognise a right of explanation. His article offers a plea in favour of the transparency of automated decision-making as a requirement for a sustainable trust in a quantified and data-driven society. As he puts it, algorithms and training data scrutiny and auditability are cornerstones of trustworthy AI.

With *Johan Rochel* and *Jean-Henry Morin* we also present a Digital Responsibility Index to quantify the responsibility of economic actors.

In his contribution, *Lexo Zardiashvili* investigates why and how to develop a responsible use of AI within police services and build a groundwork for hard regulation in the law enforcement environment of the Netherlands.

Peter J. Scott analyses historical failures of artificial intelligence and proposes a classification scheme for categorising future failures, while *James D. Miller* shows how time-inconsistency increases the challenge of building an AGI aligned with humanity's values.

Nicolas Miailhe proposes an analysis of specific use cases, to achieve Sustainable Development Goals (SDG) and formulates proposals for multi-stakeholder collaboration and new kinds of 'public-private-people' partnerships which will reconcile technical, ethical, legal, commercial, and operational frameworks. He advances new international initiatives, such as the Global Data Access Framework and the AI4SDG Center spearheaded as part of a wider international partnership called AI Commons.

Frederic Marty's contribution deals with the governance of platforms. He shows that an increasing use of AI can substantially improve performance in several areas and improve the level of trust in platforms and advanced user dissatisfaction detection tools.

Finally, *Nadisha-Marie Aliman* addresses the complexity of AI governance with safety-relevant, ethical and legal implications at an international level. She also provides novel constructive recommendations for an SDG informed AI governance and an AI-assisted approach to the SDG endeavour. AI governance could aim at a sustainable transdisciplinary scientific approach instantiated within a corrective socio-technological feedback-loop. She emphasises the need of a strong education and appropriate institutions for the realisation of this potentially robust AI governance strategy.

A Continuously Improving AI Governance Legal Framework

This Special Issue is a first contribution to the discussion of a meaningful AI governance legal framework. It proposes an overview of some of the multifaceted aspects of this topic as well as some concrete proposals to policy-makers as a way forward towards an effective human-centred AI governance. The framework that will be shaped, mainly based on bilateral and multilateral agreements between States, will require continuous improvement.

Eva Thelisson
Guest Editor
AI Transparency Institute

Power in Times of Artificial Intelligence[1]

This issue of Delphi is about power in confusing times, in times of artificial intelligence. It shows what the new technological power means for the fundamental freedoms of us humans and our democracy. A wise starting point is that AI must not be considered in isolation, but rather in the context of the concentration of economic power and digital technological power as it exists today. This is so because AI is developed and deployed to a large extent by those major digital players colloquially called the GAFAM (Google, Apple, Facebook, Amazon, Microsoft) which already have a strong grip on shaping the internet and digital technologies as we all use them. AI will be added to existing technology and business models and increase their grip even further, if we do not take the appropriate measures of regulation. The analysis of AI requires a holistic view of business models of these digital technologies and of the power they already exert today.

We need to understand not only theoretical potentials benefits of AI, which without doubt exist. We must also and foremost understand the power that is created by the combination of the different digital technologies in the hands of the corporations that dominate the internet and the state, and which, due to the rapid pace of technological development, unfolds its own dynamic that challenges democratic processes.

To understand this power and its consequences, a holistic view is needed which goes beyond market impacts. We must ask what it means for government and democracy that nearly all software for the thinking and communicating state, whether on the level of the EU or EU Member States, is procured from Microsoft and that nearly all information is stored on cloud systems. 90% of these systems are owned by US suppliers, with Amazon accounting for almost 30%. We must also be aware that more than 90% of internet searches are carried out on Google, which in turn knows more about everyone individually than individuals and their family members themselves. The fact that an ever growing section of society exclusively gets its news from Facebook and YouTube must also be a concern. What will the impact of AI, developed and deployed by the thus already powerful corporations be on individuals, democracy, governments and markets?

Technology and (economic and political) power are entering into an ever closer symbiosis. A technology that knows more about man and the world than man knows about himself, and that is given ever more decision-making powers, leads to a massive asymmetry of knowledge and power in the relationship between man and machine.

Classical models of action and decision-making in democratic societies are challenged by these developments. The question of technical power and the control of technical power is raised in a new way.

DOI: 10.21552/delphi/2019/4/4

[1] In March, Paul Nemitz and Matthias Pfeffer are publishing *Prinzip Mensch – Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz* (Dietz Verlag) <https://prinzipmenscheu.wordpress.com/>. An English edition is forthcoming later this year.
The author is writing here in his personal capacity, not necessarily representing the position of the European Commission.

Who will decide in future? And, as Shoshana Zuboff asks, 'Who decides, who decides?'[2]

When technology changes the power to shape things so radically, it is not surprising that the fundamental intellectual and cultural concepts on which modern societies are based are subjected to a stress test.

We are already experiencing the second stage of the digital 'revolution' with the current upheavals of populism, fake news, foreign propaganda and the manipulation of companies like Cambridge Analytica, based on Facebook data. And now that we look forward to AI and quantum computers, it is worth taking a look back at the beginning of the digital age to understand and learn why the great hopes of freedom and empowerment of individuals that were associated with it have largely not been fulfilled. On the contrary, we now live in a world not only of unsustainable climate change and pollution, but also of an increasingly unsustainable concentration of power and undermining of democracy and individual freedom, including informational self-determination.

In the current second phase, we can no longer afford the mistakes of the early days of digital technology and the global Internet. Technology and knowledge are developing rapidly, seemingly exploding (some speak of an exponential increase), which should lead to a transition to a whole new quality in the near future.

On the other hand, there are the deliberately slowed down processes of deliberative democracies. Slowed down, because it is an important experience of human rule, that reflective and discursive processes are vital before opinion-forming and decision-making processes in democracies are completed. A consequence of this insight is also the separation of powers and the traditional guarantees of the free press.

If technology creates facts and develops faster than democracies decide, does that mean that in this game of hare and hedgehog, technology will win for sure? Does technology even have its own developmental logic, which is proving increasingly immune to democratic control? Today, technology is creating facts at a pace that risks answering the question of power in its favour by this speed alone.

The question of who will rule in the future and who will make the decisions must be asked today in light of developments in AI and Quantum computing. We risk being ruled by AI not only through artificially intelligent systems which self – develop, as identified by Stuart Russel and others,[3] but also through the application of these technologies by powerful corporations to dominate our democracies and free will, both individual and collective.

Whoever wants to answer these questions with a firm commitment to democracy must not only bring the representatives of technology and democracy into a new conversation. We also need a clear commitment to support the good functioning of democratic process by the 'Technical Intelligentsia', a clear commitment to the rule by democracy and the rule of law rather than the rule of technological power and speed. This also means: Democracy must be willing to use its most noble tool, the law, to the set the rules in this ever more technologically colonised world, including for AI.

[2] Soshanna Zuboff, *The Age of Surveillance Capitalism* (Profile Books 2019)

[3] Stuart Russel, *Human Compatible, Artificial Intelligence and the Problem of Control* (Viking/Penguin 2019)

In his seminal Study of 1976,[4] Eugen Kogon, a frequent panellist with Adorno and Horkheimer, the protagonists of the critical Frankfurt School, showed that the political attitudes of engineers in Germany are characterised by a high degree of responsibility for the political and societal impacts of their inventions. It was the time in which 'The Physicists' by the swiss play right Dürrenmatt had been read in school by all children on their path to an entry exam for university. It is this sense of responsibility, which at the time was spurred by the threat of weapons of mass destruction and atomic power, which today must be mobilised for fending off the threats to individual freedom, fundamental rights, democracy and sustainability through unchecked technological power and its concentration in the hands of few powerful companies, at the top of the stock exchange.

Paul Nemitz
Directorate-General for Justice and Consumers
European Commission

[4] Eugen Kogon, *Die Stunde der Ingenieure* (Düsseldorf 1976)

The Right to an Explanation

An Interpretation and Defense

*Maël Pégny, Eva Thelisson and Issam Ibnouhsein**

The opacity of some recent Machine Learning (ML) techniques have raised fundamental questions on their explainability, and prompted the creation of a research subdomain, Explainable Artificial Intelligence (XAI). Opacity would be particularly problematic if those methods were used in the context of administrative decision-making, since most democratic countries grant to their citizens a right to receive an explanation of the decisions affecting them. If this demand for explanation were not satisfied, the very use of AI methods in such contexts might be called into question. In this paper, we discuss and defend the relevance of an ideal right to an explanation. It is essential both for the efficiency and accountability of decision procedures, both for public administration and private entities controlling the access to essential social goods. We answer several objections against this right, which pretend that it would be at best inefficient in practice or at worst play the role of a legal smokescreen. If those worst-case scenarios are definitely in the realm of possibilities, they are by no means an essential vice of the right to an explanation. This right should not be dismissed, but defended and further studied to increase its practical relevance.

I. Introduction

There is a fundamental ambiguity in the current use of the term *explainability* in the AI community. On the one hand, *explainability* or *(human) interpretability* denotes a fundamental scientific problem, the problem of understanding the behaviour of complex ML systems, which can lead to the development of sophisticated techniques. On the other hand, the term *explainability* is also used to denote a pedagogical problem, the problem of explaining to a lay audience, be they policy-makers or ordinary citizens, the behaviour and outcomes of those systems. Those two challenges are not completely independent: of course, a computer scientist needs to have a firm scientific grasp on a given issue before she tries to give a pedagogical explanation to a lay audience. They nevertheless need to be distinguished if we are to understand the considerable pedagogical challenges raised by ML procedures. In this paper, the terms *explanation*, *explainable* and *explainability* will have the pedagogical meaning by default. We will talk about decision explainability when the explanans will be an output that can be described as a decision.

The need for explainability is made more urgent by the use of opaque ML techniques in contexts where the public has a right to demand an explanation of the decisions affecting them.[1] The use of some of the most sophisticated ML techniques as an aid to decision-making might thus be compromised if those techniques are not explainable.

However, some authors have recently made light of the right to an explanation, dismissing it as a toothless legal tool at best, or a smokescreen giving the illusion of a right at worst. Although those are real possibilities of perversion of the right, they are by no means an essential vice: legal and technical strategies can be enforced to make it a fruitful legal tool.

In order to make this point, we will first explain the political stakes of the right to an explanation (Section II). We will then present the recent objections to this right, and show its promoters that we can as-

DOI: 10.21552/delphi/2019/4/5

* Maël Pégny, Postdoctoral Fellow, Archives Henri Poincaré, Université de Lorraine (Nancy), Membre Associé IHPST (Paris 1); Eva Thelisson, University of Fribourg, MIT Connexion Science, for correspondence: eva.thelisson@unifr.com; Issam Ibnouhsein, Quantmetry, Paris, for correspondence: iibnouhsein@quantmetry.com

[1] The persons affected by an administrative decision might of course be moral as well as physical persons. However, explanations can only be processed by human beings.

similate them to improve the conception and enforcement of this legal tool. We will first examine Veale and Edwards' objection that the right to an explanation might only provide an illusion of a right (Section III), and Floridi et al's objection against the relevance of counterfactual explanations (Section IV).

II. The Political Stakes of the Right to an Explanation

1. The Relevance of a Right to an Explanation for Government Transparency

Although most of this paper deals with generic issues of AI-assisted decision-making, it is important to stress the relevance of bureaucratic procedures. Bureaucratic procedures, public and private, are one of the main surfaces of contact between the public and systematized decision procedures, and as such they are a huge organisational and political issue.

When it comes to bureaucratic procedures in government, there is a general legal ideal of government transparency, which translates into a ‘right to an explanation’: the citizens have a right to be given an explanation of the administrative decisions affecting them. The right to an explanation we mention here is a generic philosophical ideal, not its particular and perhaps flawed implementation in any given system of positive law, eg the EU General Data Protection Regulation (GDPR) or the French *Loi sur la République numérique*. However, our discussion of this ideal right will of course be informed by the positive legal systems and the challenges raised by their application. This right to an explanation should impose explainability as a pre-condition for the use of AI systems in bureaucratic procedures.

Social contract theories clarify why government transparency is a pre-requisite to citizens' trust in such procedures.

Government transparency offers some safeguards to citizens: in its *Essay Concerning the True Original Extent and End of Civil Government*[2], John Locke argues that the Law of Nature commands that we do not harm others. In this conception, government is based on the voluntary agreements between citizens and government to care for each other. A standard of due care obliges the government to protect its citizens. If individuals consent to create a political society and a government, they receive in counterpart laws, judges to adjudicate laws, and the executive power necessary to enforce these laws. The recognition of a right to an explanation is part of a sustainability policy for a State, acting in a transparent and responsible manner, ie in reference to its duty of care. The right to an explanation is thus part of the concept of government accountability: it will facilitate the demonstration by the user of a breach of the duty of care.

The right to an explanation also plays many roles in the concrete interactions between government and citizens. It is of course the basis for appealing from a given decision. It also plays a decisive role in raising awareness of their rights and interests among citizens. It is worth being reminded that for many citizens their interaction with administrative officials is their only source of information on the legal environment affecting them. The explanation of a specific decision gives them an opportunity to improve their understanding of this environment, and conceive a strategy to defend their rights and interests. One might of course think of detrimental effects of this adaptation to the legal environment, such as the exploitation of legal loopholes for tax evasion. However, one should not reduce this strategic adaptation to those negative examples. The citizens' ability to adapt to their legal environment is often a desirable thing, which is encouraged by governments through incentives. For example, the US proposes financial and tax incentives to encourage its sustainable development policies and promote the use of energy-saving technologies.[3]

Having decisions without explanations would thus cut one of the main channels of communication between government and citizens. The debate on explainability should not fall prey to a crude opposition between procedural efficiency and respect of rights. Firstly, giving explanations creates opportunity to correct many mistakes. Secondly, procedures without explanations would lose some of their main functionalities, especially their ability to increase awareness of rights and channel incentives. Administration without explanation would not be systematically more efficient: it would be maimed.

[2] John Locke, *Two Treatises of Government* (Peter Laslett, ed, CUP 1983)

[3] US Government, Database of State Incentives for Renewables and Efficiency (2019) http://www.dsireusa.org/DSIRE accessed 30 January 2020

As a consequence, both the defenders of fundamental rights and the promoters of government efficiency should support explainability.

2. Relevance for Private Entities

Our insistence on government transparency does not mean that the right to an explanation is irrelevant for private entities. For the time being, most positive laws allow private companies to treat their procedures as trade secrets. However, as algorithms such as scoring systems used by banks, insurance companies and HR departments have a considerable impact on the general public, they should also fall, in some way or another, under the purview of an ideal right to an explanation. One could even venture to say that entities deciding who gets a loan, a job, a house or an insurance play a *de facto* governmental function, and should as such be subject to some form of accountability. The people have the right to understand the procedural environment that shapes their lives, regardless of the public or private nature of the procedural agent, or we would otherwise, to quote F. Pasquale and D. Citron's fine writing, pave the way to 'a new feudal order of unaccountable reputational intermediaries'.[4] However, articulating the legal consequences of such a viewpoint would be beyond the scope of such a short paper (for more legal reflections on the accountability of private algorithmic decision-making[5]), as it would entail a careful examination of the tension between the right to an explanation and IP rights.

However, some aspects of the right to an explanation in positive law already apply to private entities, and are worthy of comment. Amazon is facing lawsuits for this reason. An automated decision making process, without any human intervention, provides warnings and decides automatically to fire employees on the basis of input data.[6] As the productivity metrics are proprietary, employees cannot understand on which basis automated decisions are taken, despite the fact that the decisions have legal effects on the concerned person. No transparency is made on the principles and values encoded in the design of the algorithms. The result of this lawsuit may confirm the legal relevance of transparency and explanation. Based on the modernised Convention 108, employees are entitled to have knowledge on the logic involved by the algorithmic decision making process and have the right to object.[7]

In UK law, the right to an explanation might also be instrumental in extending the duty of care to private actors. In April 2019, the UK published a White Paper on Online Harms presenting statutory measures taken by the UK to reinforce the accountability of online economic actors like Facebook, Google, Snapchat, or Fortnite. The UK recognises a duty of care of online economic actors. Companies will be held to account for tackling a comprehensive set of online harms, ranging from illegal activity and content to behaviours which are harmful but not necessarily illegal. An independent regulatory body would enforce the new regulatory framework and benefit from enforcement powers. An annual Transparency Report will explain which organisational measures have been taken to avoid harming the users. In this perspective, the explanation of bureaucratic procedures and decisions can be seen as a due dilligence element and as a proof of its duty of care.[8]

In this perspective, the burden of proof lies with the company or the State. Therefore, the Online Harms White Paper is an historic paper. It obliges the digital actors, be they public or private, to be transparent and to publish an annual report bringing the proof that they behaved in a responsible manner and explaining how to the users and to the State.

III. Veale and Edwards' Objections: An Illusion of A Right?

In their comprehensive and thorough papers[9,11], Edwards and Veale have formulated several arguments

[4] Frank Pasquale and Danielle Citron, 'The Scored Society – Due Process for Automated Predictions' (2014) 89 Washington Law Review 1

[5] Frank Pasquale, *Black Box Society. The Secret Algorithms that Control Money and Information* (Harvard University Press 2015)

[6] Colin Lecher., 'How Amazon Automatically Tracks and Fires Warehouse Workers for "Productivity"',(*The Verge*, 2019) <https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations> accessed 30 January 2020

[7] Conseil de l'Europe, Convention 108 (2019) <https://www.coe.int/fr/web/data-protection/newsroom/-/asset_publisher/7oll6Oj8pbV8/content/modernisation-of-convention-108> accessed 30 January 2020

[8] UK Government, Online Harm White Paper (2019) <https://www.gov.uk/government/consultations/online-harms-white-paper> accessed 30 January 2020

[9] Lilian Edwards and Michael Veale, 'Slaves to the Algorithm?' (2017) 16 Duke Law and Technology Review 18

[11] Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions?"' (2018) 16 IEEE Security and Privacy 46

limiting the relevance of the right to an explanation, and defended the view that other approaches might be more fruitful to promote the rights of groups and individuals. They even warned against a degenerescence of that right into an illusion of a right, just as ticking the box of a User Consent Form creates an illusion of consent: 'the search for a legally enforceable right to an explanation may be at best distracting and at worst nurture a new kind of "transparency fallacy" to match the existing phenomenon of "meaningless consent"'[10].

First, the right to an explanation is not a magical, one-fits-all solution to every data and algorithm-related problem. Instead of holding that right to such an unrealistic standard, one should consider it as a necessary but insufficient condition for the protection of citizens' rights, and wonder whether one would like to live in a society where institutions are *not* required to provide explanations for their decisions. The right to an explanation should be part of a package including other approaches that are all relevant to a fair algorithmic society, such as, to name a few, the right to erasure, the right to data portability, structural due process in government agencies, auditing bodies, certification mechanisms, privacy and fairness by design.

Furthermore, the right to an explanation should not be confused with a 'duty to understand'. Individual subjects should not be burdened with an obligation to understand all the procedures affecting them, as it would represent a crushing intellectual load. It will sometimes remain a better solution to rely on a government auditing agency or a trusted expert: after all, that is what we do when we hire a legal counsel. The right to an explanation should not be read as a denial of the necessity of an intellectual division of labor and delegation of said labor, and it should not be used to burden the ordinary citizen with an intellectual workload no individual can possibly face. However, just as the complexities of positive legal systems are no excuse to make laws incomprehensible to ordinary citizens, the complexities of software are no excuse to make them incomprehensible to the people affected by them.

Finally, we agree with Veale and Edwards that the explanation of some algorithms, especially ML algorithms, will face considerable intellectual challenges, and might have some fundamental limitations.

Nevertheless, we object to a strong reading of Veale and Edwards' conclusion that would reduce the right to an explanation to an intellectual dead end, not even worthy of exploration. Our position is not rooted in *a priori* optimism on the chances of success of explanation. It is rooted in the methodological belief that such chances can not be evaluated by purely *a priori* arguments, and must be the object of a thorough empirical investigation (for an example of such an empirical investigation[12]). The real-life explainability of algorithms depends on many issues, such as open scientific questions on human interpretability of complex models, the types of questions asked or likely to be asked by the public, their relative frequency, the type of information and abstraction level adequate to answer those questions, and the type of decision-making abilities with which we want to empower the public through those explanations. This complex web of issues is worth being explored. Even if the right to an explanation were to fail as a practical endeavor, exploring the explainability of our decision procedures is a fundamental work on the intellectual division of labor, and the flow of knowledge, or lack thereof, in our social system, and it should not be given up upon. However, we believe that the work that has already been done in explainability, such as Lage, Isaac and al 2018 and[13,14] warrants the more optimistic conjecture that some relevant demands for explanation can be answered. The explainability of algorithmic procedures and the right to an explanation are not dead ends: they are vast avenues yet to be explored.

Furthermore, the right to an explanation does not only need an empirical investigation: it demands a normative reflection. In our understanding, the right to an explanation is highly normative in at least two respects. The first is that its aim is not to produce explanations that are accepted, but explanations that are honest. There is thus a necessary preliminary reflection on the nature of an honest explanation, as opposed to a rhetorical move permitting a quick-and-easy acceptance of the procedure and its results. The second is the importance of understanding not only

[10] ibid 81

[12] Isaac Lage et al, 'An Evaluation of the Human Interpretability of Explanation' (2018) NIPS Conference 32, Montréal, Canada

[13] Sandra Wachter et al, 'Counterfactual Explanation Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 Harvard Journal of Law and Technology 842

[14] Sandra Wachter et al, 'Explaining Explanations in AI' (2019) ACM FAT* 19 Conference

the questions that are asked but also the questions that should be asked. If we only focus on the questions that are currently asked, we will reproduce the current biases of administrative power, where certain populations have little knowledge or understanding of their rights, little contact with administrative institutions, or even a hostile relation with them. Those populations are unlikely to ask questions, or to ask the questions that would be truly helpful to them. It is the duty of the administration to be pro-active, to reach out to marginalised populations, and to make a normative effort to guess the kind of explanations that could truly help all of the affected individuals. This requires a deep normative reflection on the functions of the administration, its ideal relation to the population it is supposed to serve, and the role of explanation in those functions and relations.

IV. Floridi's Objections Against the Relevance of Counterfactuals for the Right to an Explanation

In a recent paper, Floridi et al[15] strongly object against a particular method to implement the right to an explanation, ie counterfactual reasoning. Counterfactual reasoning is a philosophical name for 'what would happen if...' reasoning. Answering those questions is obviously crucial to understand the role played by various factors in a decision, and to empower citizens with the ability of strategic adaptation. As such, they are a vital part of virtually any incentive policies: citizens cannot adapt their behaviour to incentives if they don't understand what would happen if they adopt the incentivised behaviour. If it would turn out that it is impossible to use counterfactual reasoning for complex algorithmic systems, then the relevance of the right to an explanation, if it would not be completely annihilated, would be drastically reduced.

It is precisely the point made by Floridi et al in their recent paper, which argues that counterfactual explanations would provide very limited interpretability to the public or the technical community. Counterfactual explanations could actually be used to generate a 'scroll-down menu' of excuses for illegal decisions: instead of admitting the use of an illegal factor *x*, such as gender, the culprit could choose among numerous, innocuous factors to provide fake explanations, such as 'your loan would have been granted if you had higher income'. Counterfactual reasoning would then be inefficient at best, and toxic at worst, and the same could be said in a large part for the right to an explanation.

First of all, this is actually a generic problem of legal explanation. A competent (and shrewd) legal expert is able to provide explanations for decisions that make them look compliant with the law, even if the actual reasons for the decisions are illegal. That is particularly problematic for individual decisions, as it is then impossible to use ordinary statistical tools to demonstrate the presence of biases.

In the case of human decisions, we do not have access to the privacy of an individual's brain: human decisions are thus by nature opaque. More often than not we do not have access to the oral deliberations of a given group, which reduces their traceability. Asking for an explanation is thus primarily a means of pressure, as are many interrogation techniques: the persons in charge of providing an explanation will have to commit to a story, which might be deemed implausible through further questioning or the discovery of new evidence. This pressure acts not only as a way to discover wrongdoings, but also as a deterrent to illegal behaviour. The power of such a tool is of course limited, and some culprits might get away with an illegal decision but again, that's a generic legal problem.

Has the situation changed with automated decision-making? It is here necessary to distinguish between cases. In the most favorable cases the situation is altered for the better, as for some computer systems we do have access to the true reasons of a decision. Automated decision systems can be probed in ways impossible for the human mind, and if the right technical conditions of human interpretability and traceability are met, we might have direct access to the true motivations of a given decision. Moreover, Floridi et al's objection seems to be founded on a scenario where an explanation could always be freely chosen from a scroll-down menu of excuses. If that is obviously in the realm of possibility, the extraction of the reasons for automated decisions might be

[15] Luciano Floridi et al, 'From What to How: An Initial Review of Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices' (*Arxiv Preprint*, 2019) <https://arxiv.org/ftp/arxiv/papers/1905/1905.06876.pdf> accessed 30 January 2020

made by a trusted third party, possibly a secure dedicated piece of software, that would actually warrant against such maneuvers. In those cases, the right to an explanation, and in particular the exploration of counterfactuals, would evolve in a direction opposite to the scenario explored in their paper in favor of more transparency and accountability. If it could be used to automate the art of excuse-making and formal compliance, automated decision-making could also be used to increase traceability and warrant the access to genuine explanations and counterfactual reasoning. The alleged defect of counterfactual reasoning is thus just a defect of a particular technological scenario which could be actively prevented.

Some other cases might of course not be as favorable. Technical conditions of traceability might not always be met, even if the legislation should encourage a positive evolution towards traceability whenever the right to an explanation applies. In some cases, the decision might not be entirely automated, which could add more degrees of freedom for a culprit to make up a false explanation. In other cases, the will to explain an actual decision or to explore counterfactual decisions might face the challenges raised by opacity, making an individual decision hard to explain even for the expert, but that would not necessarily make it easier to generate fake counterfactuals.

Furthermore, Floridi et al's argument seems to assume that for a vast majority of decisions *y*, it will be possible to find an array of factors x_o,..., x_{n-1} in order to formulate counterfactual arguments such as ‘you would have had the position if you had college education’ or ‘you would have been considered if this were a senior position’. It is the availability of such factors that make the production of excuses possible, hiding the true (and possibly illegal) decision factor x_j behind a ‘just-so’ story. However, we see no reason to assume such a possibility in the vast majority of cases, and its existence is another interesting topic for empirical investigation. Furthermore, those decision factors might have a pre-determined, hierarchised influence on a given outcome. A good explanation will also provide the user with that information, making her harder to fool with a ‘just-so’ story. For instance, she might know that her values for two factors are enough to grant her a positive decision, no matter what the other values might be. She might also know that another applicant with similar values has been accepted, making her resistant to fake explanations. It is thus impossible in the general case to pick any factor to justify any decision you wish. If counterfactual explanations are mixed with explanations of the causal relevance of each factor, as is the case in some current ‘black box’ explanation approaches (see references above), it will be much harder to generate fake explanations at will. Floridi and al's objection mistakes again the peculiarities of some scenario for an essential feature of counterfactual explanations. Combined with the relevant information, counterfactual reasoning could be a means to resist disingenuous explanations instead of a means to generate them.

V. Conclusion

The discussion of AI assisted decision making and explainability should not get stuck into a crude opposition between respect of rights and efficiency. Explanations are not a decorative feature of bureaucratic procedures: they are a major communication channel between government and its citizens, and for many of those citizens, the only channel they have. Explanations allow to increase citizens' awareness of their rights to open new opportunities, to correct mistakes and to incentivise behaviour.

It is all too easy to be dismissive with the right to an explanation. The right is by no means a sufficient warrant of a fair algorithmic society, and it could easily be perverted and emptied of its true meaning in practice. However, it is without a doubt a necessary component of a fair algorithmic society. It has to be considered in the right legal and technical context to be assessed fairly, and avoid mistaking the features of a particular scenario for essential features of this right. Dismissing the right to an explanation in the discussion of algorithmic fairness would be a terrible mistake, as it would leave out of public sight a right which needs careful interpretation, vigorous enforcement and dedicated technical work to bear its fruits.

AI Governance: Digital Responsibility as a Building Block

Towards an Index of Digital Responsibility

*Eva Thelisson, Jean-Henry Morin and Johan Rochel**

The rapid development of AI-based technologies significantly impacts almost all human activities as they are tied to already existing underlying systems and services. In order to make sure that these technologies are at least transparent if not provably beneficial for human beings and society and represent a true progress, AI governance will play a key role. In this paper, we propose to reflect on the notion of 'digital responsibility' to account for the responsibility of economic actors. Our objective is to provide an outline of what digital responsibility is and to propose a Digital Responsibility Index to assess corporate behavior. We argue that a Digital Responsibility Index can play a central role in restoring trust in a data-driven economy and create a virtuous circle, contributing to a sustainable growth. This perspective is part of AI governance because it provides a concrete way of quantifying the implementation of AI principles in corporate practice.

I. Introduction

AI technologies now underlie almost all systems and services used in transforming how we live, learn, work, engage, vote, socialise, travel, help, etc. AI technologies suffer from a lack of transparency, which raises the question of how liability risks will be taken into consideration by policy-makers.[1] In addition, companies leveraging the underlying data are building empires concentrating power over people to a level never achieved before.[2] This becomes all the more significant as a few actors dominating the market hold the data of billions of people, potentially influencing their lives and practices. A prominent example is Facebook, which in barely ten years rose from being an internal college dating site to the biggest global social network service ever built with almost one third of the world population being registered and sharing their most intimate information.[3]

Corporate responsibility is a concept larger and older than that applied in the digital field. The UNCTAD (United Nations Conference on Trade and Development) has discussed for more than 25 years the Social Responsibility of Transnational Corporations with all stakeholders involved (Governments, Corporations, Civil Society, etc.). This paper will focus on digital responsibility only. This is a very important concept in ever more digitalized societies. In both cases we have similar dimensions: politics, ethics, legal issues, human rights, finance, geopolitics, etc.

AI plays an important role in the digitalization of our societies. This transition is an ongoing process we need to cope with and organise amongst different stakeholders to achieve a balance preventing one stakeholder from dominating the others at their expense. Broadly, we identify three major stakeholders: private companies (industry), public authorities (state) and individuals (society).[7] Implementing

DOI: 10.21552/delphi/2019/4/6

* Eva Thelisson, Massachusetts Institute of Technology, MIT Connexion Science Lab, Boston, USA, Co-Founder of the AI Transparency Institute. For correspondence: <evathelisson@protonmail.com>.
Jean-Henry Morin, Institute of Information Service Science (ISS), Centre Universitaire d'Informatique (CUI), Geneva School of Social Sciences, University of Geneva, Geneva, Switzerland. For correspondence: <jhmorin@unige.ch>.
Johan Rochel, Faculty of Law, University of Zürich; ethix - Lab for Innovation Ethics, Zurich, Switzerland. For correspondence: <johan.rochel@gmail.com>.

[1] Iria Giuffrida, 'Liability for AI Decision-Making: Some Legal and Ethical Considerations' (2019) 88 Fordham L Rev 439

[2] James E Bessen, 'The Policy Challenge of Artificial Intelligence' (2018) CPI Antitrust Chronicle, Boston Univ School of Law, Law and Economics Research Paper No 18-16

[3] Roger McNamee, *Zucked: Waking Up to the Facebook Catastrophe* (Harper Collins 2019)

checks and balances which enable economic growth and innovation, while fostering the respect of democratic and human values and principles is the core purpose of designing what is called 'AI governance'.[4]

In this governance scheme, we need strong legal and regulatory frameworks. Ex-ante mechanisms might concern pre-authorisation by safety authorities for high risk products and services. Ex-post mechanisms pertain to safeguards (eg privacy and human rights impact assessments) and effective remedies (eg class actions) to protect individuals' rights enabling a sustainable digital society. The EU has given a strong signal in this direction with the General Data Protection Regulation reform (GDPR) basically applying to all states as long as the data concerns EU data subjects. The EU Guidelines on AI Trustworthiness are another example of this political will to cooperate at the EU level and to engage in a sustainable and responsible way in the use of artificial intelligence. At the OECD level, the Principles on Artificial Intelligence set standards for AI and promote artificial intelligence that is innovative and trustworthy and that respects human rights and democratic values. In June 2019, the G20 adopted human-centered AI Principles that draw from the OECD AI Principle. The IEEE is also developing ethical standards mainly for intelligent and autonomous systems.

As part of this AI governance scheme, we also need to consider private companies as duty-bearers. But how should private companies define their responsibility? We argue for the need to introduce 'digital responsibility' as a criteria to approach the responsibility of companies with respect to digital matters, and AI in particular. We will show how this concept might be used as an operationalising concept for corporate responsibility to contribute to a sustainable and human-centered digital society. This approach – focused on the responsibility of corporations – is the way we think we will have the best chance to contribute to the issue as it is complementary to developing strong legal and regulatory frameworks.

This contribution is organised as follows. First, we will present and discuss the concept of responsibility and their bearers with a particular focus on the corporate sector. Secondly, we will introduce digital responsibility and propose a way of classifying its components. Thirdly, we will outline a tentative design of a digital responsibility index and describe how it could be used in a way to help organisations both assess where they stand and help them progress on a path towards improving their digital responsibility.

II. Responsibility in General and its Relation to the Digital Realm

In order to define 'digital responsibility' we must first understand responsibility in general before addressing the specific features of digital responsibility. In a nutshell, the concept of digital responsibility represents a specific category of the concept of responsibility used in moral philosophy.

We focus on a specific type of responsibility holder, namely companies.[5] In the company-focused literature, digital responsibility is often understood as 'corporate digital responsibility', meaning a kind of digital Corporate Social Responsibility (CSR).[6] We complement this literature by taking a broader view, highlighting the different fundamental meanings of responsibility before specifying what these meanings could mean in the digital realm.[7] In doing so, we put our approach in the context of the literature on responsible innovation.[8] This approach appears

[4] Urs Gasser and Virgilio AF Almeida, 'A Layered Model for AI Governance' (2017) 21 IEEE Internet Computing 6, 58-62; Allan Dafoe, 'AI Governance: A Research Agenda' (2018) Future of Humanity Institute, University of Oxford; Alan F T Winfield and Marina Jirotka, 'Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence Systems' (2018) 376 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2133, 20180085

[5] For a similar approach, Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers* (Springer 2017)

[6] For a conceptualisation on CSR in digital time, focusing on 'new ways of communicating existing issues and new responsibilities associated with the corporate use of digital technologies', Georgiana Grigore et al, 'New Corporate Responsibilities in the Digital Economy ' in A Theofilou, G Grigore and A Stancu (eds), *Corporate Social Responsibility in the Post-Financial Crisis Era* (Palgrave Macmillan 2017) 43; See also C Thorun, 'Corporate Digital Responsibility: Unternehmerische Verantwortung in der digitalen Welt' in Gärtner C and Heinrich C (eds), *Fallstudien zur Digitalen Transformation* (Springer Gabler 2018)

[7] See also S Pellé and B Reber, 'Responsible Innovation in the Light of Moral Responsibility' (2015) 15 Journal on Chain and Network Science 107, 111

[8] On this literature, Job Timmermans and Vincent Blok, 'A Critical Hermeneutic Reflection on the Paradigm-Level Assumptions Underlying Responsible innovation' (2018) Synthese Vincent Blok and Pieter Lemmens, 'The Emerging Concept of Responsible Innovation. Three Reasons Why it is Questionable and Calls for a Radical Transformation of the Concept of Innovation' in Bert-Jaap Koops (ed), *Responsible Innovation: Concepts, Approaches, and Applications* (Springer 2015); On the distinct understandings of responsibility at stake in innovation, Ibo van de Poel and Martin Sand, 'Varieties of Responsibility: Two Problems of Responsible Innovation' (2018) Synthese 1

to be conceptually more solid and offers a comprehensive overview of the implications of responsibility in the digital realm.

1. Responsibility in General

Responsibility is one of the most fundamental issues addressed in moral and political philosophy.[9] The concept has given rise to abundant literature in diverse philosophical traditions. We focus on two important aspects: distinguishing between distinct meanings of responsibility, and briefly addressing the question of which entity could bear responsibility.

In short, the main idea of responsibility could be summarised in the following way: when a person or an entity performs or fails to perform a morally significant action, we think that a particular kind of response is warranted. This relation between what has been done or should be done and the specific response is what we grasp with the idea of responsibility.

This idea of responsibility only makes sense if one essential condition is fulfilled. This condition refers to the freedom a person or an entity needs to have when performing a specific action. This condition might be further specified as having the capacity to act differently. If there was no other option for how to act, the question of responsibility cannot be raised in the same way. It shall be made clear that the 'free will' question about determinism looms large in this debate.[10] We can add to this freedom-condition a knowledge-condition. In specific situations, the quality of knowledge available when making a specific decision might have a fundamental impact when it comes to assessing whether an agent's responsibility is engaged. In situations where it was impossible to know the kind of harm (which would be) done, we need to adjust the kind of responsibility at stake.

Of course, this question of having the capacity to act differently, or the benchmark used to assess whether enough knowledge was available is ultimately a social question. It depends on the context in which the situation occurs. AI systems make the debate more complex. These elements emphasise that our understanding of responsibility is always defined in a specific context. A reaction - or absence of reaction - in turn further defines this context. As put by Eshleman, 'through the reactive attitudes (eg resentment) we communicate to fellow members of the moral community our interpersonal expectations for a reasonable degree of goodwill.[11]

We need to briefly address the issue of the type of agent that could be said to bear responsibility. If we assume that a human being might bear this type of responsibility, the question is more complicated for a company. Broadly, two conditions need to be fulfilled.

The first one pertains to the identification of a company being able to carry out an action. This condition helps to distinguish an organised company from the mere aggregation of individuals. It could be described and measured by observing internal decision-making procedures or representation mechanisms (executives, board members, etc.). The second one pertains to the required quality of the decision taken by the company. It mirrors and qualifies the 'freedom' condition mentioned above for the case of human beings. The decisions taken by the company must show a certain degree of rationality. It must be able to pursue something and to take reasons into account.

In the context of this paper, we assume these two conditions are fulfilled in the case of standard companies. To give an example, these conditions might be arguable in the case of a decentralised autonomous organization on a blockchain. It is not clear whether a fully decentralised organization fulfills the conditions required to be able to bear responsibility.[12]

2. Two Understandings of Responsibility

Assuming that the condition of freedom is fulfilled, we can distinguish between two types of responsibility: negative and positive responsibilities.[13]

[9] For an overview, Andrew Eshleman, 'Moral Responsibility' (2014) Stanford Encyclopedia of Philosophy

[10] For references, Matthew Talbert, 'Moral Responsibility' (2019) Stanford Encyclopedia of Philosophy 1.

[11] Eshleman, 'Moral Responsibility' (2014) Stanford Encyclopedia of Philosophy 2.2

[12] For a legal perspective on this issue, Daniel Kraus, Thierry Obrist and Olivier Hari, Blockchains, Smart contracts, Decentralised Autonomous Organisations and the Law (Edward Elgar Publishing 2019)

[13] For a similar distinction, Pellé and Reber, 'Responsible Innovation in the Light of Moral Responsibility' (2015) 15 Journal on Chain and Network Science 107 111

Negative responsibility identifies acts or omissions which should not be carried out/should not have been carried out. In a nutshell, it is about preventing harm. This understanding of responsibility is linked to concepts such as blameworthiness, liability or accountability. It legitimates fair compensation in order to repair damage a posteriori and to punish the originator of the negligence or the fault.[14] It is used for both individuals and companies. Negative responsibility puts a clear focus on the causal role played by an agent in carrying out an action. It is often used in the context of identifying past wrongdoings. By anticipating future reactions linked to a specific act or omission, negative responsibility could be used to assess decisions made in the present (see below the further distinction between prospective and retrospective responsibility).

When referring to this negative responsibility, we need to address the following questions:

- Which value(s) are used to determine the types of harm which generate responsibility?
- Which kind of benchmark is used to assess one's contribution to this harm?
- How should the type and extent of compensation be determined?

The positive dimension of responsibility shares a main insight of the negative dimension: it links the behavior of an agent to a particular situation in the world. However, unlike the situation described above, it focuses on a morally relevant situation that is not the result of the company's action. It puts the focus on the capacity of an agent (freedom and knowledge) to pursue a specific course of action for the sake of addressing a morally relevant situation. When we ascribe positive responsibility to an entity, we do not tell a causal story about the entity. Instead, we specify what this entity should be doing in the world. As put by Smiley, positive responsibility is used to distribute moral labor for future decisions.[15]

A good example of this positive responsibility is inspired by Peter Singer's 'child drowning' thought experiment.[16] While jogging in the park, you notice a child drowning in a pond. It is completely safe for you to step into the water and take the child out of it. In this example, your responsibility to act appears to be fully engaged. By taking action, you might prevent a morally disastrous situation (the child's death), without taking major risks for yourself. The situation here is absolutely clear: you are the only one able to help the child. The distribution of responsibility upon diverse agents is not an issue here. Similarly, the moral urgency of the situation is indisputable. A number of agents might be responsible for contributing to solving a specific problem. Furthermore, the assessment of the problem at stake might be itself disputed.

When referring to this positive responsibility, we need to address the following questions:

- What is/are the moral value(s) used to describe the morally relevant situation at stake?
- If the moral value(s) collide(s) with another one, what are their relative priorities?
- If diverse agents should address this morally relevant situation, what is one agent's fair share?
- Who played a causal role for the creation of the morally relevant situation in the first place?
- Do particular practical elements impact on an agent's responsibility, such as a specific capacity to address the situation or detrimental conditions (eg costs)?

Determining an agent's positive responsibility should be understood as an ongoing process.[17] As for the negative responsibility, this process is impacted by and does itself impact the social context in which it takes place. This is most clearly the case with the assessment of the moral value of the situation at stake and the definition of what is seen as a 'problem'.

a. Prospective and Retrospective Responsibility

Another concept of theoretical value is the distinction between *retrospective* responsibility and *prospective* responsibility. Both dimensions clearly apply to the realm of digital responsibility. While retrospective responsibility has to do with the question which responsibility an actor bears for an action (or omission) in the past, prospective responsibility is

[14] Poel and Sand, 'Varieties of Responsibility: Two Problems of Responsible Innovation' (2018) Synthese 1

[15] Marion Smiley, 'Collective Responsibility' (2017) Stanford Encyclopedia of Philosophy § 7

[16] Peter Singer, *Practical Ethics* (Cambridge University Press 1980)

[17] Pellé and Reber, 'Responsible Innovation in the Light of Moral Responsibility' (2015) 15 Journal on Chain and Network Science 107 113

Table 1: Responsibility Conceptual Framework. Source: Authors' elaboration

	Negative responsibility	Positive responsibility
Retrospective responsibility	Address problems created by the company in the past => repair and compensate	Address problems which the company had not created, but which represented a moral urgency => contribute to reducing the harmful consequences of the actions of others
Prospective responsibility	Address problems which the company could create now and in the future => prevent	Address problems which the company does not focus its actions on, but which might represent a moral urgency, eg with regard to the environment in which the company acts => contribute to preventing the harmful consequences of the actions of others

about actions to be taken (or to be omitted) in the future. Adding these two dimensions, a conceptual framework of the notion of responsibility can be presented as in Table 1.

b. Digital Responsibility in Particular

These two understandings of responsibility might be applied in the digital realm, representing what we will call ‘digital responsibility’. Here again, some distinctions are useful in structuring the debate. Firstly, responsibility might be used to account for classical issues of business ethics in the digital economy. The issues are well known and are ‘simply’ found in a different setting.[18] In these cases, digital companies need to address similar criticisms as other types of companies. The fact that they develop and sell digital technologies does not prevent them from being caught up in questionable choices regarding for instance taxes, bribing, corruption, the behaviors of their employees. In this sense, digital responsibility, understood as the responsibility of digital companies, is the same as the responsibility which other types of companies have.

Secondly, and more importantly for this paper, new uses and actions made possible by digital technologies might create new responsibilities. This issue is of crucial importance as companies are taking note of the growing awareness of their clients, and more broadly of the public. What is required is to structure the different fields of this new digital responsibility and investigate which values are at stake.

III. Decomposing Digital Responsibility in its Constituencies

If the company’s responsibility is our main focus, we need to distinguish between the distinct target categories of actors which the company is responsible for (its ‘constituencies’).[19] Each category may have different or sometimes even opposing goals depending on the ethical conflicts, potentially leading to conflicting outcomes. Let us consider two basic initial categories, defined along their relations towards the company.

Digital services and products: this category considers all actors directly or indirectly affected by the digital services and products developed by a company.

- *Customers/users:* This category includes individuals who use a digital service/product. They have a direct interest in digital responsibility, for example in terms of privacy and data protection, or the responsible design of systems and services, etc.
- *Society:* In this category, the actions/omissions of a company may negatively impact societal and economic qualities of life (eg cloud computing strategy).

[18] Grigore et al, 'New Corporate Responsibilities in the Digital Economy ' in Theofilou, Grigore and Stancu (eds), Corporate Social Responsibility in the Post-Financial Crisis Era (Palgrave Macmillan 2017) 49

[19] For a similar reflection, Klaus-Dieter Altmeppen et al, 'Öffentlichkeit, Verantwortung und Gemeinwohl im digitalen Zeitalter' (2019) 64 Publizistik 59, 67

- *Governance of the company:* This category considers all individuals or entities directly affected by the digitization of the company. It means that it also applies to companies which do not produce digital services or products.
- *Employees:* This category includes individuals in a situation of employment within the company. They have an indirect interest in the outcome of digital responsibility in terms of training and skills, quality of workplace, value alignment with the management etc.
- *Shareholders/owners:* This category includes individuals or entities holding a share or who are the owners of the company. Their focus may cover issues of governance, technology watch, planning for skills, revenue, etc.
- *Suppliers and subcontractors:* This category includes all third party commercial entities working with the company to deliver its products and services. Data access and joint liability with the data controller are key concerns.

We will now consider the different thematic dimensions of digital responsibility and apply each of these dimensions to the different constituencies. Following this, these dimensions are matched with the constituencies, from the perspective of both negative and positive responsibility (see Figure 1).

1. Securing Autonomy and Privacy

Artificial intelligence systems should be designed, implemented and brought to the market respecting the values of autonomy and privacy. The crucial capacity of human beings to freely take decisions and act according to them should be respected. This capacity requires a protected personal sphere, as detailed in the General Data Protection Regulation and the Convention 108 of the Council of Europe and its protocol.

2. Respecting Equality

For companies active in a democratic and liberal environment, the value of equality among individuals is crucial. This value represents an ideal for society: every individual should be recognised as equally important human beings. This equality is understood here as a basic moral equality among all human beings. This raises the key question for every policy regarding equality: which features of being human do we want to protect from being grounds for discrimination? Some grounds for differential treatment are legitimate, while others are not considered as such. From this very fundamental understanding of equality, we might formulate and justify more specific demands (economic equality, equality of opportunities, etc.). Algorithmic tools raise new challenges regarding social justice and equality, due to the key role of datasets quality. If the datasets are not representative of the population of a country, then the product or service based on the training data may not work for some categories of persons and be harmful for them (eg a cancer diagnosis system not working for black people).[20]

3. Dealing with Data

Digital technologies make possible, contribute to and take advantage of the 'dataification' of the world. Almost every aspect of our individual and collective lives might be expressed and documented in the form of data. Respecting privacy and autonomy requires consent-based and proportionate data collection, storage, use and transfer. Consent should be informed and explicit, based on full information as to the type, scope and purpose of the data being collected. Dealing with these data should respect the principle of good faith and due care.[21]

4. Dealing with Algorithms

Digital responsibility deals with the challenges raised by the wide use of algorithms in different settings. Across all these settings, digital responsibility calls for the use of algorithms which respect safety, autonomy and, more generally, the principles linked to the rule of law. This means in particular the capacity to reconstruct and explain decisions taken by algorithms. It also means the precautionary use of algorithms in settings especially sensitive for autonomy.

[20] Adam Conner-Simons and Rachel Gordon, 'Using AI to Predict Breast Cancer and Personalize Care' (2019) MIT News

[21] *In latin, bonus pater familias.*

Furthermore, digital responsibility bears upon the wide use of algorithm in automating different tasks within the company. Even if the AI system is not a decision-making system but only assists human beings in making decisions, there is a risk that professionals rely too much on data analytics which raises a *de facto* delegation of responsibility to the system (eg radiology).

5. Taking Impact on the Environment into Account

Digital technologies have an important impact on resources and, more broadly, on the environment.[22] While this impact might be positive (eg digital technologies reducing the general consumption of resources, such as in smart city projects), the use of these technologies rely upon resources such as electricity, space, water to cool down datacenters, but also on specific materials used in the production of hardware (and the recycling thereof).

6. Ensuring a Fair Transition

Digital technologies bring changes which impact individuals and society. This impact might be positive, but it might also be negative. Companies have a responsibility to identify, accompany these changes and to proactively contribute to a successful transition enabling a fair and sustainable digital society. One solution could be to allocate a share to data subjects as reward for the data collected and monetised. This solution would align the interests of shareholders and data subjects, whose data are an asset for companies.

IV. The Digital Responsibility Index: Tentative Design and Potential Use

In Table 2 (see Annex), we present these different dimensions along the different constituencies and the distinction between negative and positive responsibility. Taken together, they form the core of the Digital Responsibility Index. We argue that companies should use this Index as both an assessment and an improvement metric. Through a self-assessment approach, companies can assess their level of maturity and eventually engage in an improvement process on the basis of the Index. In the absence of recognised formal legal frameworks, a soft-compliance approach may be appealing and could even become a business advantage in an age where customers are increasingly putting sustainability and responsibility pressure on companies on digital issues. Self-assessment methods combined with approaches using maturity based models can be considered as valuable for the company in improving their digital readiness. Overall, the Index helps better understand the issues and assess where companies stand, but also to provide them with an improvement process should they decide to increase their maturity level.

Concretely, our proposition is to formulate each of the normative desiderata entailed by the Index in the form of a question which the company should ask itself. Each question receives a weight as well as a criticality indicator which represents the importance of the question with regards to the overall topic. With the support of an evaluation system, we deduce a summary dashboard presenting the strengths and weaknesses of a company policy. We deduce from this dashboard some charts presenting the state of maturity of the company or of a research project compared to a predefined threshold (see Figure 1). We also deduce a global scoring, aggregating the digital responsibility constituents into a percentage index for example to show the maturity level. These can be organised into categories depending on the focus. This would allow for a graphical representation in the form of a radar clearly showing the coverage of the company. We also combine this approach with a simple maturity model recommending what is needed to engage and progress to the next level of digital responsibility.

While such an approach may be very useful for private companies, it may also help in shaping the debate on digital responsibility of organisations prior technology transfers of AI-based systems in the market. Just as with social responsibility, rating agencies or analysts specialised in investment recommendations may use the same criteria to ask the tough questions. This also helps to show large digital actors that society is more aware and sensitive to these is-

[22] M Stuermer, G Abu-Tayeh and T Myrach, 'Digital Sustainability: Basic Conditions for Sustainable Digital Artifacts and Their Ecosystems' (2017) 12 Sustainability Science 247

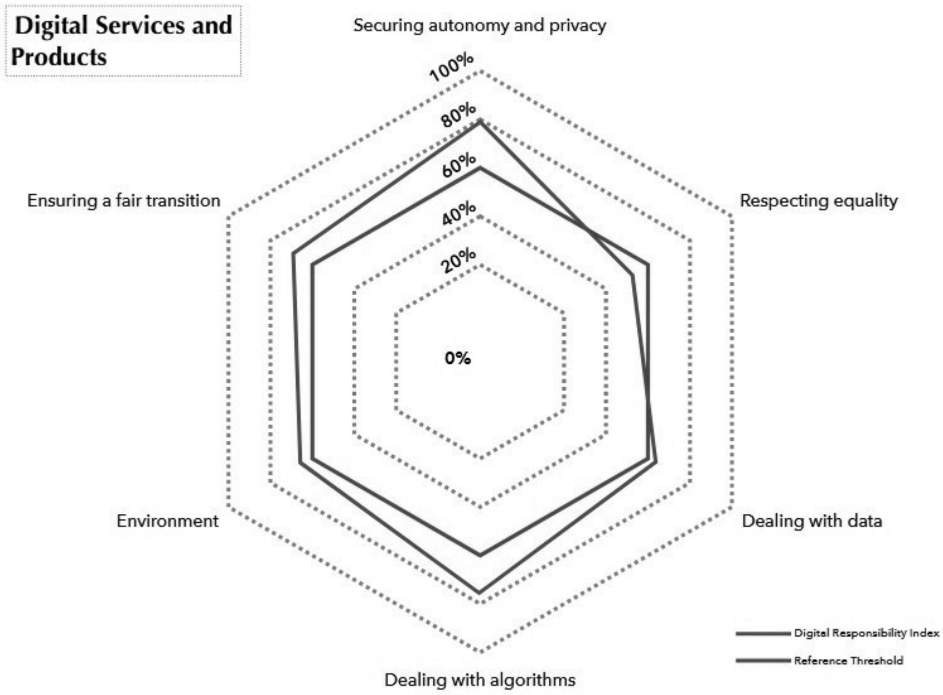


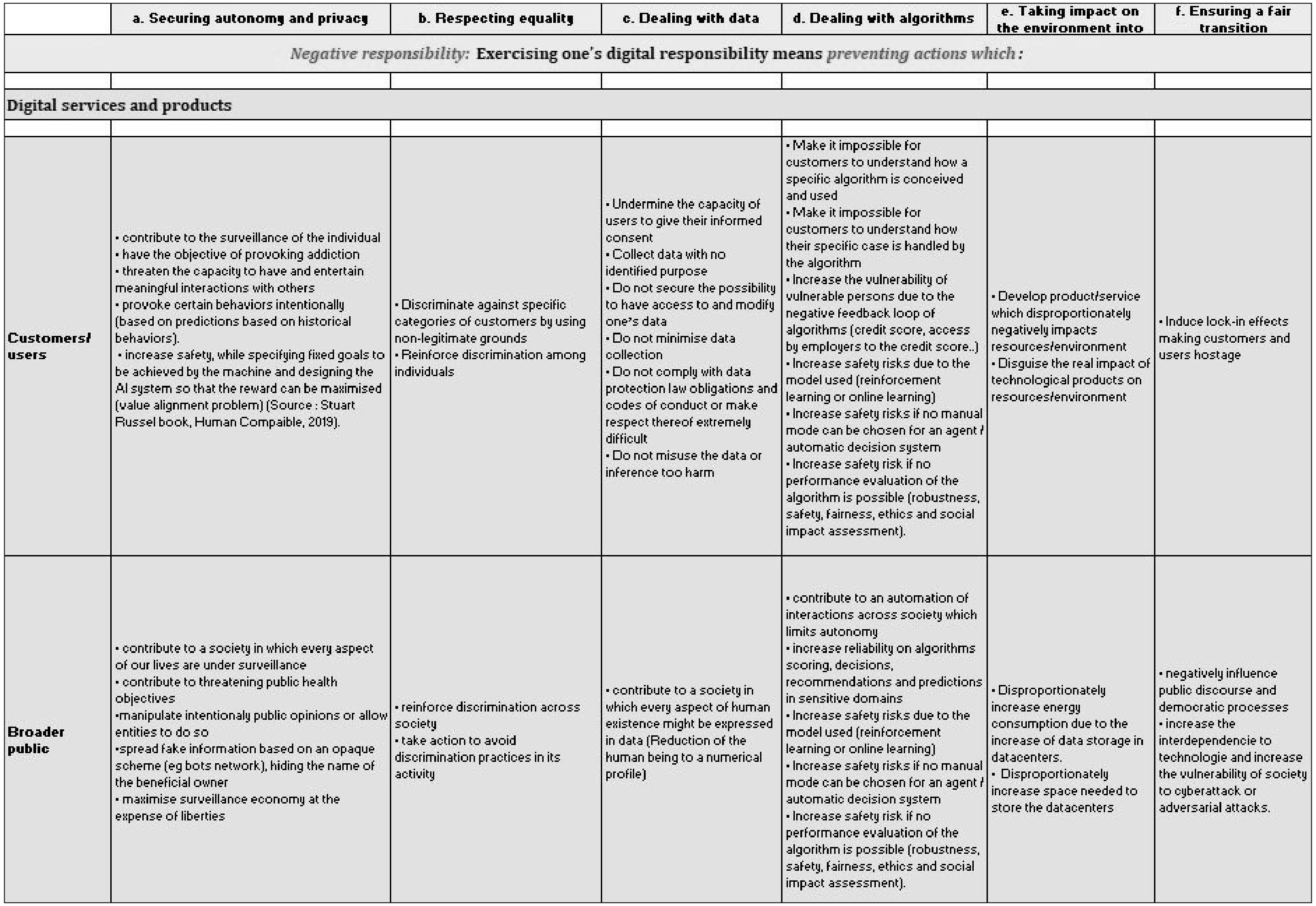
Figure 1: Digital Responsibility Index Radar
Source: AI Transparency Institute

sues in a similar way in which social responsibility made its way into the corporate environment over time.

V. Conclusion

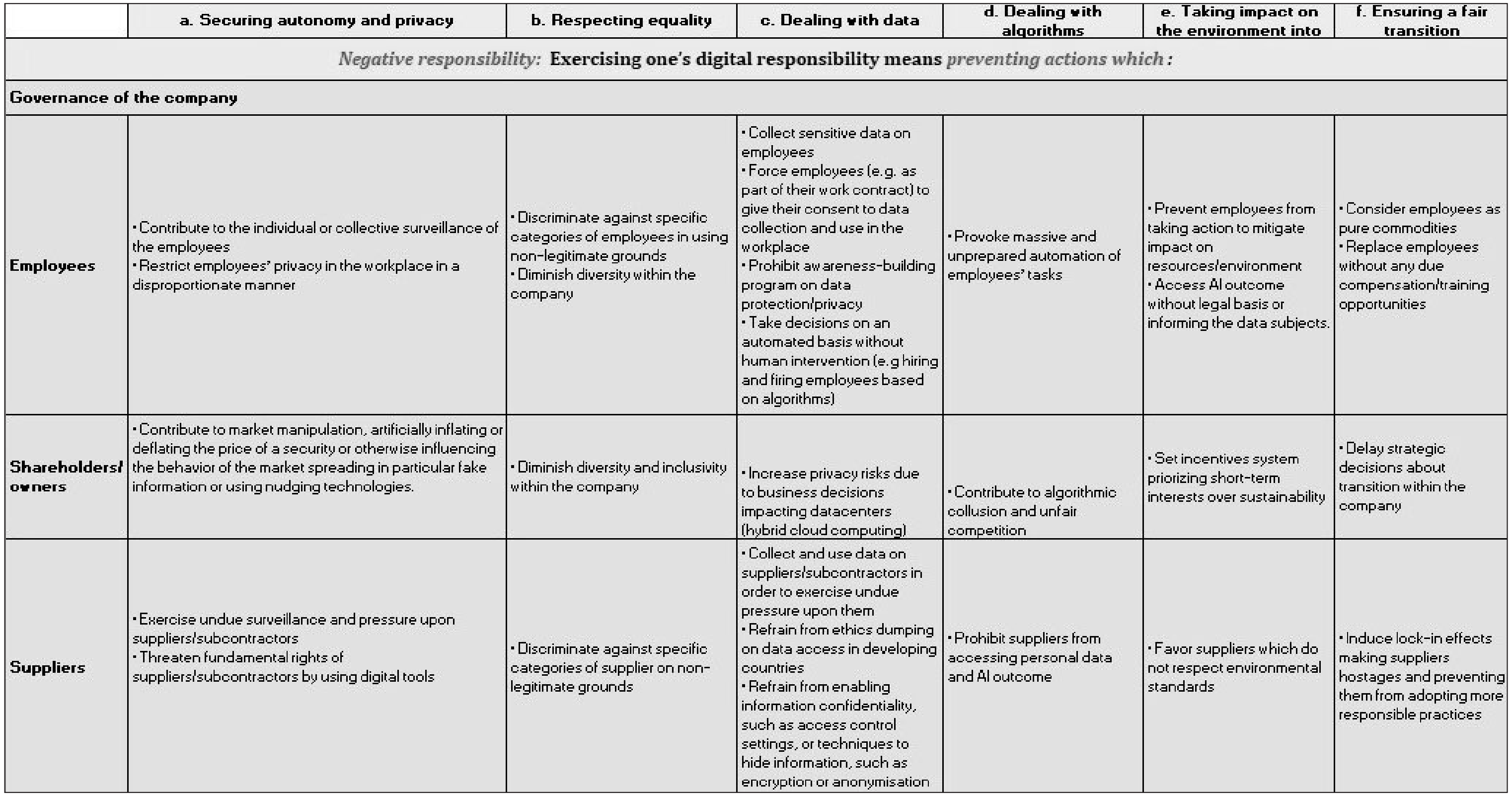
The framing of an AI governance scheme is a question for the State(s) (in a legal and regulatory way), but it is also a challenge for private companies acting within the broader normative framework of a free market economy. In order to address their responsibility, we have outlined the concept of a 'digital responsibility' and developed a Digital Responsibility Index. This Index brings together a key distinction between negative and positive responsibility, the identification of constituencies and the thematic dimensions of digital responsibility. All together, they form the Index. This Index might be used as a self-assessment tool for private companies as well as an evaluation framework for large corporations. As we also think it is important to allow improvement, we propose the use of maturity models to help progress along the various levels to ultimately help reach a level of being a trustworthy and 'digitally responsible company'.

Annex



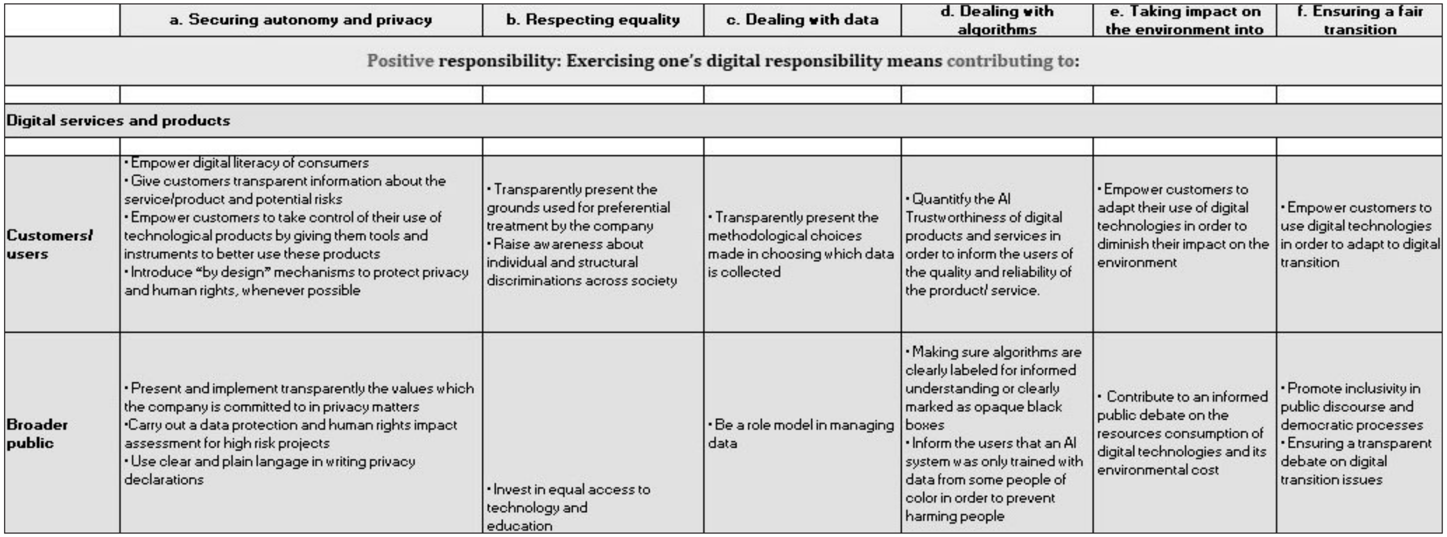
	a. Securing autonomy and privacy	b. Respecting equality	c. Dealing with data	d. Dealing with algorithms	e. Taking impact on the environment into	f. Ensuring a fair transition
Negative responsibility: **Exercising one's digital responsibility means** *preventing actions which:*						
Digital services and products						
Customers/ users	• contribute to the surveillance of the individual • have the objective of provoking addiction • threaten the capacity to have and entertain meaningful interactions with others • provoke certain behaviors intentionally (based on predictions based on historical behaviors). • increase safety, while specifying fixed goals to be achieved by the machine and designing the AI system so that the reward can be maximised (value alignment problem) (Source : Stuart Russel book, Human Compaible, 2019).	• Discriminate against specific categories of customers by using non-legitimate grounds • Reinforce discrimination among individuals	• Undermine the capacity of users to give their informed consent • Collect data with no identified purpose • Do not secure the possibility to have access to and modify one's data • Do not minimise data collection • Do not comply with data protection law obligations and codes of conduct or make respect thereof extremely difficult • Do not misuse the data or inference too harm	• Make it impossible for customers to understand how a specific algorithm is conceived and used • Make it impossible for customers to understand how their specific case is handled by the algorithm • Increase the vulnerability of vulnerable persons due to the negative feedback loop of algorithms (credit score, access by employers to the credit score..) • Increase safety risks due to the model used (reinforcement learning or online learning) • Increase safety risks if no manual mode can be chosen for an agent / automatic decision system • Increase safety risk if no performance evaluation of the algorithm is possible (robustness, safety, fairness, ethics and social impact assessment).	• Develop product/service which disproportionately negatively impacts resources/environment • Disguise the real impact of technological products on resources/environment	• Induce lock-in effects making customers and users hostage
Broader public	• contribute to a society in which every aspect of our lives are under surveillance • contribute to threatening public health objectives •manipulate intentionally public opinions or allow entities to do so •spread fake information based on an opaque scheme (eg bots network), hiding the name of the beneficial owner • maximise surveillance economy at the expense of liberties	• reinforce discrimination across society • take action to avoid discrimination practices in its activity	• contribute to a society in which every aspect of human existence might be expressed in data (Reduction of the human being to a numerical profile)	• contribute to an automation of interactions across society which limits autonomy • increase reliability on algorithms scoring, decisions, recommendations and predictions in sensitive domains • Increase safety risks due to the model used (reinforcement learning or online learning) • Increase safety risks if no manual mode can be chosen for an agent / automatic decision system • Increase safety risk if no performance evaluation of the algorithm is possible (robustness, safety, fairness, ethics and social impact assessment).	• Disproportionately increase energy consumption due to the increase of data storage in datacenters. • Disproportionately increase space needed to store the datacenters	• negatively influence public discourse and democratic processes • increase the interdependencie to technologie and increase the vulnerability of society to cyberattack or adversarial attacks.

Table 2: Ethos Matrix. Source: Authors' elaboration



	a. Securing autonomy and privacy	b. Respecting equality	c. Dealing with data	d. Dealing with algorithms	e. Taking impact on the environment into	f. Ensuring a fair transition
Negative responsibility: **Exercising one's digital responsibility means** *preventing actions which:*						
Governance of the company						
Employees	• Contribute to the individual or collective surveillance of the employees • Restrict employees' privacy in the workplace in a disproportionate manner	• Discriminate against specific categories of employees in using non-legitimate grounds • Diminish diversity within the company	• Collect sensitive data on employees • Force employees (e.g. as part of their work contract) to give their consent to data collection and use in the workplace • Prohibit awareness-building program on data protection/privacy • Take decisions on an automated basis without human intervention (e.g hiring and firing employees based on algorithms)	• Provoke massive and unprepared automation of employees' tasks	• Prevent employees from taking action to mitigate impact on resources/environment • Access AI outcome without legal basis or informing the data subjects.	• Consider employees as pure commodities • Replace employees without any due compensation/training opportunities
Shareholders/ owners	• Contribute to market manipulation, artificially inflating or deflating the price of a security or otherwise influencing the behavior of the market spreading in particular fake information or using nudging technologies.	• Diminish diversity and inclusivity within the company	• Increase privacy risks due to business decisions impacting datacenters (hybrid cloud computing)	• Contribute to algorithmic collusion and unfair competition	• Set incentives system priorizing short-term interests over sustainability	• Delay strategic decisions about transition within the company
Suppliers	• Exercise undue surveillance and pressure upon suppliers/subcontractors • Threaten fundamental rights of suppliers/subcontractors by using digital tools	• Discriminate against specific categories of supplier on non-legitimate grounds	• Collect and use data on suppliers/subcontractors in order to exercise undue pressure upon them • Refrain from ethics dumping on data access in developing countries • Refrain from enabling information confidentiality, such as access control settings, or techniques to hide information, such as encryption or anonymisation	• Prohibit suppliers from accessing personal data and AI outcome	• Favor suppliers which do not respect environmental standards	• Induce lock-in effects making suppliers hostages and preventing them from adopting more responsible practices

Continuation of Table 2: Ethos Matrix



	a. Securing autonomy and privacy	b. Respecting equality	c. Dealing with data	d. Dealing with algorithms	e. Taking impact on the environment into	f. Ensuring a fair transition
Positive responsibility: Exercising one's digital responsibility means contributing to:						
Digital services and products						
Customers/ users	• Empower digital literacy of consumers • Give customers transparent information about the service/product and potential risks • Empower customers to take control of their use of technological products by giving them tools and instruments to better use these products • Introduce "by design" mechanisms to protect privacy and human rights, whenever possible	• Transparently present the grounds used for preferential treatment by the company • Raise awareness about individual and structural discriminations across society	• Transparently present the methodological choices made in choosing which data is collected	• Quantify the AI Trustworthiness of digital products and services in order to inform the users of the quality and reliability of the product/ service.	• Empower customers to adapt their use of digital technologies in order to diminish their impact on the environment	• Empower customers to use digital technologies in order to adapt to digital transition
Broader public	• Present and implement transparently the values which the company is committed to in privacy matters •Carry out a data protection and human rights impact assessment for high risk projects • Use clear and plain langage in writing privacy declarations	• Invest in equal access to technology and education	• Be a role model in managing data	• Making sure algorithms are clearly labeled for informed understanding or clearly marked as opaque black boxes • Inform the users that an AI system was only trained with data from some people of color in order to prevent harming people	• Contribute to an informed public debate on the resources consumption of digital technologies and its environmental cost	• Promote inclusivity in public discourse and democratic processes • Ensuring a transparent debate on digital transition issues

Continuation of Table 2: Ethos Matrix

	a. Securing autonomy and privacy	b. Respecting equality	c. Dealing with data	d. Dealing with algorithms	e. Taking impact on the environment into	f. Ensuring a fair transition
Positive responsibility: Exercising one's digital responsibility means contributing to:						
Governance of the company						
Employees	• Use digital tools to empower employees to be active members of the company	• Use digital tools to raise awareness and address discrimination patterns among employees	• Be transparent about the data collected about employees and its management	• Train employees in Digital Technologies and foster autonomy • Verify the correctness of data which is crucial with AI systems which are based on historical data.	• Reward employees' contribution to better environmental impact	• Proactively identify evolutions of the skills required for employees and the required educational effort • Plan required actions to ensure that employees are able to keep pace with technological developments, most importantly from the perspective of their skills
Shareholders/ owners	• Publish concrete actions carried out to promote responsible and safe digital gouvernance	• Reinforce diversity and inclusivity within the company	• Promote trust and ethics in dealing with data use and management • Refrain from Ethics dumping on data in developing countries	• Invest in prioritizing trustworthy algorithmic decision making systems, products and services.	• Only invest in environmentally friendly companies • Assess the cost of environmental impact of activities • Implement ethical standards and give the priority to suppliers implementing ethical standards	• Invest in quality processes and internal controls to verify the fairness of algorithms and digital technologies • Invest in quality labels and certification mechanisms • Invest in companies having received such a certification, seal or labels on digital technologies
Suppliers	• Promote transparency on data processing and security • Communicate pledges and concrete corresponding results, maybe in verifiable ways	• Implement best practices on equality and promote actively this concept internally	• Refrain from Ethics dumping on data in developing countries (I don't know what this sentence is supposed to mean) • Proactively communicate with suppliers about their rights and obligations in data law	• Proactively communicate with suppliers about the way algorithms are used to design collaborations and conduct specific projects	• Empower suppliers to diminish impact on resources	• Proactively communicate with suppliers about technological evolutions impacting their activities

Continuation of Table 2: Ethos Matrix

AI Ethics for Law Enforcement

A Study into Requirements for Responsible Use of AI at the Dutch Police

*Lexo Zardiashvili, Jordi Bieger, Francien Dechesne and Virginia Dignum**

This article analyses the findings of empirical research to identify possible consequences of using Artificial Intelligence (AI) for and by the police in the Netherlands, and ethical dimensions involved. We list the morally salient requirements the police need to adhere to for ensuring the responsible use of AI and, further, analyse the role of such requirements for governance of AI in the law enforcement domain. We list the essential research questions that can, on the one hand, help to flesh out more detailed criteria for the responsible use of AI in the police, and on the other, build a groundwork for the hard-regulation in the law enforcement environment of the Netherlands.

I. Introduction

Under the Dutch Police Law (Politiewet 2012) the task of the Dutch police is two-fold: (1) to ensure maintaining the rule of law and (2) to provide assistance to those in need.[1] The police have a special role in society that involves a constitutional right to use violence for the enforcement of the law.[2] For the police to function and realise its objectives, society has to deem the police as legitimate and trust that it is effective in its tasks.[3] In order for the police to be trustworthy in their *efficacy*, they must continuously innovate to evolve with developments, stay ahead of criminals' new strategies and capabilities, and utilise new methods and technology for the fulfilment of their tasks.[4] In order for the police to be trustworthy in their *use of power*, the police must demonstrate goodwill and respect for the rights of civilians. The National Police greatly values the trust of Dutch citizens, which was measured to be the highest of any measured institution in 2017.[5] It is important to retain this trust, also when introducing new technologies such as Artificial Intelligence (AI) that have a fundamental impact on the nature of their operations and interactions with society.[6]

AI has many potentially beneficial applications in law enforcement including predictive policing, automated monitoring, (pre-) processing large amounts of data (eg, image recognition from confiscated digital devices, police reports or digitized cold cases), finding case-relevant information to aid investigation and prosecution, providing more user-friendly services for civilians (eg with interactive forms or chatbots), and generally enhancing productivity and

DOI: 10.21552/delphi/2019/4/7

[*] Lexo Zardiashvili, LLM, PhD Candidate at the Center for Law and Digital Technologies Leiden Law School, Leiden University. For correspondence: a.zardiashvili@law.leidenuniv.nl. Jordi Bieger, MSc, Researcher/Teacher at the Faculty of Technology, Policy and Management, Delft University of Technology, and PhD Candidate at the Center for Analysis and Design of Intelligent Agents, Reykjavik University. For correspondence: <J.E.Bieger@tudelft.nl>. Francien Dechesne, Assistant Professor at the Center for Law and Digital Technologies, Leiden Law School, Leiden University. For correspondence: <f.dechesne@law.leidenuniv.nl>. Virginia Dignum, Associate Professor at the Faculty of Technology, Policy and Management, Delft University of Technology. For correspondence: <M.V.Dignum@tudelft.nl>.

[1] The Dutch Police Law (Politiewet) 2012

[2] Joris Boumans, 'Technologische Evoluties in Wetshandhaving en Legitimiteit: Tussen Optimisme en Onbehagen' (MSc thesis, Tilburg University 2018)

[3] Kees van der Vijver, 'Legitimiteit, gezag en politie. Een verkenning van de hedendaagse dynamiek' in C. D. van der Vijver and F. Vlek (eds), *De legitimiteit van de politie onder druk? Beschouwingen over grondslagen en ontwikkelingen van legitimiteit en legitimiteitstoekenning* (Elsevier 2006), 15-133

[4] ibid

[5] Centraal Bureau voor de Statistiek, 'Meer vertrouwen in elkaar en instituties' (*Centraal Bureau voor de Statistiek* 28 May 2018) <www.cbs.nl/nl-nl/nieuws/2018/22/meer-vertrouwen-in-elkaar-en-instituties> accessed 24 September 2019

[6] Boumans (n2)

paperless workflows. AI can be used to promote core societal values central to police operations (human dignity, freedom, equality, solidarity, democracy, and the rule of law), but, on the other hand, values carefully guarded in existing operations and procedures may also be challenged by the use of AI.

Currently the police in the Netherlands have been using AI in all applications mentioned above. For example, the 'Crime Anticipation System' (CAS) is an internally developed predictive-policing tool that aims to predict crimes with statistics based on data from various sources.[7] 'Pro-Kid 12- SI' (pronounced "Pro-Kid twelve-minus") is a rule-based system for risk assessment on children aged between 0-12 years, used nationwide by the police to prevent children from being involved in a crime or anti-social behaviour.[8] The Online Fraud Report Intake System uses NLP techniques, computational argumentation (legal informatics) and reinforcement learning to assist civilians in reporting the crime.

It is impossible to anticipate all the effects of the use of AI in society, and more specifically, in the law enforcement domain. Therefore, it is essential that adoption and use of any application be continuously evaluated, in order for the Dutch police to ensure policing practices in line with the values acknowledged by the Dutch state and the European Union.

With this goal in mind, we conducted an empirical study to identify possible consequences of using AI for, and by law enforcement and the ethical issues this may lead to. On the basis of this research, we have co-written a white paper for the Dutch police: *'AI & Ethics at the Police: Towards Responsible Use of Artificial Intelligence in the Dutch Police' (*hereafter Whitepaper).[9] It describes the state-of-the-art in AI, how it could benefit law enforcement, and what ethical concerns will need to be addressed in the use of AI in order to safeguard the legitimacy of and trust in the national police.

II. On the Law and Ethics: The Role of Ethics in Law Enforcement

Similar to other authorities of the state, the police necessarily operate within a specific legal framework. This framework includes but is not limited to preventing misuse of powers, conflicts of interest and discrimination, and is ensured through active accountability measures. The police organisation in the Netherlands is committed to protect fundamental human rights and to ensure respect for the rule of law.[10] The police is directly obliged to comply with domestic and international legal instruments that specify this commitment, like the national constitution, the EU Charter, specific national legislative acts, and the EU directives and regulations like the General Data Protection Regulation (GDPR) or Law Enforcement Directive (LED). These legal requirements apply to all police work regardless of the means used and thus include the use of AI.

In a democratic state such as the Netherlands, compliance with holding laws and regulations must be seen as a given for any application of AI. However, the application of AI raises some challenges that are not—or it is unclear if they are—covered by current legal provisions. For example, while the legislation might not require full openness, the opacity of reasoning that is inherent to some AI techniques might decrease transparency and weaken human agency in the police's decision-making, and thereby pose a threat to the legitimacy of and trust in the police.[11] Therefore, for such spaces left open by the law, the police *can*, and we advise that they *should*, incorporate 'ethics' through practical measures to ensure responsible use of AI and contribute towards enhancing (rather than limiting) legitimacy of and trust in the police.

In common use, the term 'ethics' refers to a set of accepted principles on what is (morally) right or wrong within and for a certain community. The Dutch government and the law enforcement in particular are expected to act coherently and out of the

[7] Serena Oosterloo and Gerwin van Schie, 'The Politics and Biases of the 'Crime Anticipation System' of the Dutch Police', Jo Bates, Paul D. Clough, Robert Jäschke and Jahna Otterbacher (eds), *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems* (CEUR Workshop Proceedings 2018) 30-41

[8] Karolina La Fors-Owczynik and Govert Valkenburg, 'Risk Identities: Constructing Actionable Problems in Dutch Youth', I. van der Ploeg and J. Pridmore (eds), *Digitizing Identities. Doing Identity in a Networked World* (Routledge/Taylor & Francis Group 2016) 103-124

[9] Francien Dechesne, Virginia Dignum, Lexo Zardiashvili and Jordi Bieger, 'AI and Ethics at the Police: Towards Responsible Use of Artificial Intelligence at the Dutch Police' (*Whitepaper*, 2019) https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/ai-and-ethics-at-the-police-towards-responsible-use-of-artificial-intelligence-at-the-dutch-police-2019..pdf accessed 24 September 2019

[10] Politiewet 2012 (n1), art 2

[11] Dechesne and others (n9)

principles of the Dutch (and larger European) community. This expectation of responsibility extends to the use of AI by the Dutch police. To act responsibly means to accept moral integrity and authenticity as ideals and to deploy reasonable effort toward achieving them.[12] For the Dutch government striving for moral integrity means adhering to the values of *freedom*, *equality*, and *solidarity*.[13] These values are three from four values the European Union (EU) is aiming to uphold, with *dignity* being the fourth.[14] Note that, although the Dutch government has not yet accepted proposals by a specially established commission (established by the Cabinet for constitutional amendments), to include value of *human dignity* explicitly in the text of the Dutch Constitution, it acknowledges dignity as a fundamental value that human rights aim to uphold.[15] *Human rights*, on the other hand, together with *democracy,* and *rule of law,* are often referred as the general principles of the Dutch constitution,[16] of the EU,[17] and of also larger European community (Council of Europe).[18]

The four *values* (dignity, freedom, equality, solidarity) and three *principles* (human rights, democracy, rule of law) provide a framework for the moral integrity that the Dutch government (and in this case the Dutch police) has to continuously strive towards. However, societal order as a moral milieu cannot be sustained by reference only to generally expressed values – therefore formal (statutory and case) law is intended to fill in the gap and operationalise these abstract ideals. On the other hand, such moral milieu cannot be built upon strict textually-rooted rules alone.[19] For example, in the context of state-of-the-art technology, formal law fails to be the omnibus governance solution: existing legislation is not perfectly suited to address unprecedented scope of actions that AI allows, and regulatory intervention (among other things) might prevent potential advantages from materialising.[20]

Therefore, maintaining responsible action (moral integrity) requires a proper balance to be struck between 'rule' and 'value'. What this means in the context of using AI is that, unprecedented *modus operandi* to the formal law does not relieve the Dutch police from an obligation to strive towards moral integrity. We have evaluated the use of AI by the law enforcement through the lens of the (European) *values* (dignity, freedom, equality, solidarity) and *principles* (human rights, democracy, rule of law) that the Dutch police aims to uphold, and identified *requirements* for ensuring responsible use of AI within the police.[21] We provide the overview of identified requirements in the next chapter.

III. Requirements for the Responsible Use of AI by the Dutch Police

We identified requirements and recommendations for the responsible use of AI at the Dutch police. They include, (i) accountability, (ii) transparency, (iii) privacy and data protection, (iv) fairness and inclusivity, (v) human autonomy and agency, and (vi) sociotechnical robustness and safety.[22] While these requirements are morally salient, they do not occupy the same level of hierarchy as the *values* and the *principles* discussed in the chapter II (hence the term *requirements*). Rather these requirements are intended to provide guidance on how to ensure that the police use of AI is coherent to the high-level *values* (ie dignity) and the *principles* (ie democracy):

1. *Accountability* – In the context of using AI for and by the police, 'accountability' is a requirement that refers to the ability to hold the police personnel or the entire police organisation answerable and/or

[12] Ronald Dworkin, 'Justice for Hedgehogs' (The Belknap Press, 2011) 111

[13] Ministry of Social Affairs and Employment, 'Core Values of Dutch Society' *(Pro Demos, House of Democracy and Constitution*, 2014) https://www.prodemos.nl/wp-content/uploads/2016/04/KERNWAARDEN-ENGELS-S73-623800.pdf accessed 17 October 2019

[14] Charter of Fundamental Rights of the European Union (The EU Charter), 26 October 2012, 2012/C 326/02

[15] Jan-Peter Loof, 'Human Dignity in the Netherlands' in Paolo Becchi, Klaus Mathis and Jan-Peter Loof (eds.), *Handbook of Human Dignity in Europe* (Springer International Publishing 2017) 423

[16] ibid

[17] The EU Charter, Preamble; see also European Union, ' Goals and values of the EU' https://europa.eu/european-union/about-eu/eu-in-brief_en accessed 17 October 2019

[18] Council of Europe, 'Values – Human Rights, Democracy, Rule of Law' https://www.coe.int/en/web/ about-us/values accessed 17 October 2019

[19] Chief Justice Allsop AO, 'Values in Law: How They Influence and Shape Rules and the Applications of Law' (*Hochelaga Lecture*, 2016) https://www.fedcourt.gov.au/digital-law-library/judges-speeches/chief-justice-allsop/allsop-cj-20161020#_ftn3 accessed 17 October 2019

[20] Ronald Leenes and others, 'Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues' (2017) 9 (1) Law, Innovation and Technology, 7

[21] Dechesne and others (n9)

[22] ibid

responsible (and/or sometimes liable) for an action, choice or decision by AI. Tracing (causal) responsibility can be complicated when human decision makers are (partially) replaced or augmented by AI systems that cannot themselves carry moral responsibility or be accountable. Accountability can be improved if these systems can be reviewed (auditability), and if the decisions that they make explained and justified (explainability) on the technical level. Moreover, independent evaluations should be able to verify and reproduce the AI-system's behavior in all situations (reproducibility).[23] In cases where tracing responsibility is not feasible (and possibly others), clear agreements should be made about who is accountable (eg the owner, operator or programmer of an AI system).

2. *Transparency* – Transparency is an important component in ensuring trust and figuring out who or what is accountable for potential problems with AI systems. With transparency, we must always ask 1) about what, 2) to whom and 3) how much transparency should be provided, and of course to what end. We can be transparent for example about people, rationale, operations, or data involved in decision-making. We can be transparent for courts, police organisation, or to the public. Perhaps giving everyone full access to everything is not productive, and it can even be dangerous if it lets bad actors find ways to exploit or circumvent the police's AI. Transparency is a gradual matter, and the same holds for explainability and interpretability: we have to take into account that in the context of AI only parts of a decision may be interpretable, or that explanations only give a rough idea of what happened.
3. *Privacy and Data Protection* – The Police has a (legal) obligation to take the privacy of civilians into consideration in their operations. Where civilians can reasonably expect to be private is being altered by the current technology that allows personal data from many different spheres to be processed on an unprecedented scale, also for law enforcement purposes (eg prevention, investigation, detection or prosecution of criminal offences). AI can increase the information-gathering capabilities of the police, because of its ability to combine and analyze vast quantities of data from different sources, and therefore has an immense impact on privacy.
4. *Fairness and Inclusivity* – AI systems can play an important role in the inclusivity and accessibility of police services. For instance, reporting of a crime will be accessible to more people if more reporting methods are available, eg in person at a police station, by phone and online. Intelligent chatbots can make reporting crimes more accessible for some by increasing accessibility, user friendliness and catching errors that might otherwise be made on static forms. One should however be careful that the range of methods offered is indeed usable by all, including eg blind people or (computer) illiterate people. If this is not feasible for the main method, alternatives should (continue to) be provided. AI can also increase usability by eg adding speech recognition functionality (which can help people who can't type text). It is also important to ensure that decisions informed by AI are free from bias which could result in the unfair or discriminatory treatment of (groups of) civilians. This requires rigorous acquisition, management, development and evaluation of AI systems and algorithms as well as the data they use. Since there are different conceptions of fairness, presenting different tradeoffs depending on the situation, an informed case-by-case analysis in necessary for the responsible use of AI by the police. In the end, (human) police employees will need to decide what to do with the information and recommendations provided by AI, raising questions about what kind of action is appropriate: eg if a suspect has not done anything wrong yet, but an (imperfect) AI system predicts that they might in the future, what interventions balance the rights of the as-of-yet innocent civilian with the need to prevent serious crimes?
5. *Human Autonomy and Agency* – Preserving the human sense of agency is mainly an individual-level requirement to realise the high-level values (i.c. freedom) and should help with both job satisfaction and the ability to provide meaningful human control. Problems can occur with decision support systems that recommend a course of action that must then be evaluated by a human operator. People are increasingly willing and expected to dele-

[23] Matthew Hudson, 'Artificial Intelligence Faces Reproducibility Crisis' (2018), 359 (6377) Science 725-726

gate decisions and actions to machines (eg recommender systems, search engines, navigation systems, virtual coaches and personal assistants). A possible consequence of working with AI systems is the loss of a sense of agency: the ability to act freely. Especially with systems that are very accurate in some respect, human operators may be 'nudged' to act upon the outcome of the system without further critical deliberation. This can not only invalidate an operator's sense of agency, but also fails to utilise human capabilities that AI systems typically still lack, such as commonsense reasoning, looking at the bigger picture, and adapting to unforeseen situations.

6. *(Socio-technical) Robustness and Safety* – AI systems must be developed and deployed with an awareness of the risks and benefits of their use, and an assumption that despite ample preventative measures, errors will occur. They must be *robust* to errors and/or inconsistencies in their design, development, deployment and use phases, and degrade gracefully in extraordinary situations, including adversarial interactions with malicious actors. Errors and malfunctions should be prevented as much as possible, and processes should be in place to cope with them and minimise their impact.[24] An explicit and well-formed development and evaluation process is necessary to ensure performance, robustness, security and safety.

The Dutch Police acts to maintain societal order by enforcing the law. The law itself is a set of binding rules that aim to uphold the values within society. While a set of binding rules can guide the only limited amount of police actions, societal values are always present, and the activities of the police are responsible only when adhering to these values. If AI is to be utilised, the police is compelled to take into consideration morally salient requirements described in this chapter, to ensure responsible action (responsible use of AI). How can these requirements influence the set of binding rules will be discussed in the next chapter.

IV. Ethics and the Re-evaluation of Law

Alongside the rapid development of AI, there is a proliferation of articles and policy documents about the governance of AI, some of which seem to suggest 'ethics' as the solution for ensuring responsible use of AI. Few months before we delivered the Whitepaper to the Dutch police, researchers at Berkman Klein Center identified and positioned thirty-two sets of policy documents side by side, enabling comparison between efforts from governments, companies, advocacy groups, and multi-stakeholder initiatives.[25] Thirteen of the thirty-two documents presented in this study discuss the responsibility of governments in the context of AI, as we did in our Whitepaper. These documents acknowledge that the existing set of legal rules is not able to fully deal with the impacts of AI, and propose guidance for maintaining moral integrity of governmental actions by reflecting upon ethical *values* and *principles*.[26]

However, contrary to some of these governmental[27] and most of the private sector[28] policy documents, our whitepaper did not intend to come up with the new set of *principles* for the use of AI within the Dutch police. Rather, we looked at the *values* and the *principles* that the Dutch police, as the law enforcement body of the Dutch state, is already obliged to adhere to and identified what is *required* to ensure such coherence (and therefore responsible use of AI). Moreover, we believe that ethical *values* and *laws* are 'expressions along a gradation of particularity' rather than 'clearly identifiable separate vehicles'.[29] In this sense, *law* conforms to *ethics*, as the latter provides 'a gauge to the law's flexibility', and its 'avenue for growth'.[30]

In other words, while ethical reflections provide advantages as an open norm-setting venues for the governance of AI within the law enforcement, such considerations could do more by going beyond tech-

[24] High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI'(High-Level Expert Group On Artificial Intelligence, The European Commission 2019)

[25] Jessica Fjeld and others, 'Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches' (Berkman Klein Center 2019) https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf accessed 24 September 2019

[26] *see* Federal Government of Germany, 'AI Strategy' (2019)

[27] see Smart Dubai, 'AI Principles and Ethics' (2019) https://www.smartdubai.ae/ accessed 18 October 2019

[28] see Sundar Pichai, 'AI at Google: Our Principles' (*Google*, 2018) https://www.blog.google/technology/ai/ai-principles/ accessed 18 October 2019

[29] Chief Justice Allsop AO (n 21)

[30] ibid

nical interpretations of morally salient requirements (ie accountability, transparency)[31], and serve as the lens through which existing legal frameworks (including frameworks regulating the activities of the police) are re-evaluated, to see if improvements are possible.[32] In the end, such re-evaluation seems to be the last logical step as the absence of adequate formal rules, might 'confound law by a drift into a formless void of sentiment and intuition'.[33]

V. Further Research in Responsible Use of AI in Law Enforcement

As the complete picture of the effects of the use of AI technology cannot be anticipated, not all ethical and societal impacts of the use of AI at the law enforcement body of the Netherlands could be covered in the short study of the Whitepaper.[34] Therefore, ethical evaluation of the use of AI by the law enforcement needs to be continuous to be able to transform concerns into better laws. With this goal in mind, we identified the following research directions on AI and ethics at the police,[35] divided into tracks for (1) impact on humans, (2) organisational embedding, and (3) technical work:

1. Impacts on Humans:

   a. *Impacts on Human Dignity* – Human dignity is the inviolable value upon which the human rights framework rests. It illustrates the fundamental belief in the intrinsic worth of a human being, protecting his/her autonomy and self-determination. Belief in human dignity can be understood as the raison d'être for the law the police aims to enforce.
   b. *Public Trust* – Public perception of the legitimacy of the police and subsequent trust is as important as the legal framework in which the police operate. While automation and prediction to some extent increase efficacy of the police, the study could explore if such increase in potency is desirable from the societal perspective.

2. Impacts on the Police Organisation:

   a. *Ethics Guidelines and Oversight* – The police does not operate in isolation, and the use of AI takes place across the entire judicial chain: OM, local government, the Ministry of Justice and Security, judiciary. Responsible use of AI within the Dutch police ideally follows from a robust ethics framework for the entire chain. Such a framework can establish criteria to follow throughout the AI development and application cycle.
   b. *Impacts on Police Personnel* – AI can be used to support the police organisation in achieving its goals of efficiency, traceability, uniformity and integrity. However, the change of operations may come with displacement of employees and changing roles. Research is required to ensure that workers with non-traditional skillsets fit into the police organisation in a way that empowers police personnel.

3. Technical Aspects

   a. *Explainable AI* – The aforementioned oversight can only be adequate and meaningful if automated decisions can be explained and justified on the technical level.
   b. *Justifiable/Verifiable AI* – Justification provides the reasons behind the results and the choices for particular approaches. Mathematical tools for formal verification make AI systems themselves and their decisions reviewable.

Further research is essential so that the police continues to realise their dual goals of increasing (a) efficacy and efficiency, and (b) trust and trustworthiness (to boost public trust and the perception of the legitimacy of the police). The research in the areas described above will help us re-evaluate the formal rules regarding law enforcement, and also make societal requirements transparent to both the police

[31] Corinne Cath, 'Governing artificial intelligence: ethical, legal and technical opportunities and challenges' (2018), 376 (2133) Philosophical Transactions of the Royal Society A Mathematical, Physical and Engineering Sciences

[32] Luciano Floridi, and others, 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018), 28(4) Minds and Machines 689–707

[33] Chief Justice Allsop AO (n 21)

[34] Whitepaper (n 13)

[35] Francien Dechesne, Virginia Dignum, Lexo Zardiashvili and Jordi Bieger, 'Long-Term Research Strategy for AI and Ethics at the Police' (Report 2019) https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/artificiele-intelligentie-en-ethiek-bij-de-politie/research-strategy-ai-ethics-dutch-police-final.pdf accessed 24 September 2019

and the public and ultimately enable codification in the legal frameworks.

VI. Conclusions

This article has analysed the role of the morally salient requirements for governance of AI, that were found in an empirical study within the law enforcement domain – in particular: at the Dutch Police. We have argued that there are instances, where the need for soft regulatory instrument arises, and we have described how ethical considerations can help fulfil this need. Our analysis suggests that the responsible use of AI at the Dutch police requires primarily the following requirements: accountability, transparency, privacy, fairness and inclusivity, human autonomy and agency and socio-technical robustness and safety.

Furthermore, we explored the role of these requirements in a future re-evaluation of the formal binding instruments. Finally, we identified the areas where further research is advisable for ensuring the responsible use of AI at the Dutch police. On the one hand, such research can help flesh out more detailed criteria for the police on how to adhere to the values and principles of the Dutch state. On the other, it can build a groundwork for the hard-regulation for the use of AI in the law enforcement ecosystem of the Netherlands.

Classification Schemas for Artificial Intelligence Failures

*Peter J. Scott and Roman V. Yampolskiy**

In this paper we examine historical failures of artificial intelligence (AI) and propose a classification scheme for categorising future failures. By doing so we hope that (a) the responses to future failures can be improved through applying a systematic classification that can be used to simplify the choice of response and (b) future failures can be reduced through augmenting development lifecycles with targeted risk assessments.

I. Introduction

Artificial intelligence (AI) is estimated to have a $4-6 trillion market value[1] and employ 22,000 PhD researchers.[2] It is estimated to create 133 million new roles by 2022 but to displace 75 million jobs in the same period.[3] Projections for the eventual impact of AI on humanity range from utopia[4] to extinction.[5] In many respects AI development outpaces the efforts of prognosticators to predict its progress and is inherently unpredictable.[6]

Yet all AI development is (so far) undertaken by humans, and the field of software development is noteworthy for unreliability of delivering on promises: over two-thirds of companies are more likely than not to fail in their IT projects.[7] As much effort as has been put into the discipline of software safety, it still has far to go.

Against this background of rampant failures we must evaluate the future of a technology that could evolve to human-like capabilities, usually known as *artificial general intelligence* (AGI). The spectacular advances in computing made possible by the exponential hardware improvements due to Moore's Law[8] balanced against the unknown required breakthroughs in machine cognition make predictions of AGI notoriously contentious. Estimates of how long we have before AGI will be developed range over such widely varying timelines[9] that researchers have taken to meta-analysis of the predictions through correlation against metrics such as coding experience of the predictors.[10]

Less contentious is the assertion that the development of AGI will inevitably lead to the development of ASI: *artificial superintelligence*, an AI many times more intelligent than the smartest human, if only by virtue of being able to think many times faster than a human.[11] Analysis of the approach of confining a superintelligence has concluded this would be diffi-

DOI: 10.21552/delphi/2019/4/8

* Peter J. Scott, Next Wave Institute, USA. For correspondence: <peter@humancusp.com>; Roman V. Yampolskiy, University of Louisville, Kentucky, USA, <roman.yampolskiy@louisville.edu>

[1] McKinsey Global Institute, 'Notes from the AI Frontier' <https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/notes-from-the-ai-frontier-insights-from-hundreds-of-use-cases-discussion-paper.ashx> accessed 1 November 2019

[2] Jeremy Kahn, 'Just How Shallow is the Artificial Intelligence Talent Pool?' (*Bloomberg,* 7 February 2018) <https://www.bloomberg.com/news/articles/2018-02-07/just-how-shallow-is-the-artificial-intelligence-talent-pool> accessed 1 November 2019

[3] World Economic Forum, 'The Future of Jobs Report' <http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf> accessed 1 November 2019

[4] Raymond Kurzweil, *The Singularity Is Near: When Humans Transcend Biology (*Viking 2005) 487

[5] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies (*Oxford University Press 2005)

[6] Roman Yampolskiy, 'Unpredictability of AI' (2019) *arXiv:1905.13053v1 [cs.AI]*

[7] Keith Ellis, 'The Impact Of Business Requirements On The Success Of Technology Projects' (*BA Times*, 15 February 2008) <https://www.batimes.com/articles/the-impact-of-business-requirements-on-the-success-of-technology-projects.html> accessed 1 November 2019

[8] Chris Mack, 'Fifty Years of Moore's Law' (2011) 24 IEEE Transactions on Semiconductor Manufacturing 202-207

[9] Kaj Sotala and Roman Yampolskiy, 'Corrigendum: Responses to Catastrophic AGI Risk: A Survey' (2015) 90 *Phys. Scr. 018001*

[10] Brian Tomasik, 'Predictions of AGI Takeoff Speed vs. Years Worked in Commercial Software' in *Essays on Reducing Suffering* (2014) <https://reducing-suffering.org/predictions-agi-takeoff-speed-vs-years-worked-commercial-software/> accessed 1 November 2019

[11] Vernor Vinge, 'The Coming Technological Singularity: How to Survive in the Post-human Era' in National Aeronautics and Space Administration (eds), *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace (1993)*

cult[12] if not impossible.[13] Many of the problems presented by a superintelligence resemble exercises in international diplomacy more than computer software challenges; for instance, the *value alignment problem*[14] (described therein as the 'value loading problem') of aligning AI values with humans'.

II. Definitions

We present some operational definitions of terms used in this paper.

Artificial intelligence is a shifting term whose definition is frequently debated. Its scope changes depending upon the era: during an 'AI Winter'[15] many fewer vendors are willing to identify their products as AI than during the current period of myriad AI technologies clogging the 'peak of inflated expectations' in the Gartner Hype Cycle.[16]

Failure is defined as 'the nonperformance or inability of the system or component to perform its expected function for a specified time under specified environmental conditions.'[17] This definition of failure as an event distinguishes it from an *error,* which is a static condition (or state) that may lead to a failure.

Cybersecurity has been defined as 'the organisation and collection of resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign *de jure* from *de facto* property rights.'[18] AI Safety has been defined as an extreme subset of cybersecurity: 'The goal of cybersecurity is to reduce the number of successful attacks on the system; the goal of AI Safety is to make sure zero attacks succeed in bypassing the safety mechanisms.'[19]

Intelligence definitions converge toward the idea that it '(...) measures an agent's ability to achieve goals in a wide range of environments.'[20] We do not present this definition with any intention of defining AI by applying the 'artificial' modifier to this one. Rather, this definition will be used to judge whether a software failure is instructive in the extent to which it was applying (accidentally or intentionally) intelligence in even the narrowest sense, since such application could extend to a more powerful AI.

III. AI Failure Classification

We will describe a tag schema for classifying AI failures. It is precisely because of the volatile definition of AI that we must cast a wide net in what we use for examples of AI failures, because what is classified as AI today will likely be given a less glamorous title (like 'machine vision') once it becomes commonplace. As AI pioneer John McCarthy put it, 'As soon as it works, no one calls it AI any more.'[21] Where some of our examples, therefore, may appear to be indistinguishable from failures of software that has no particular claim to the label of artificial intelligence, they are included because they are close enough to AI on the software spectrum as to be indicative of potential failure modes of AI.

1. Historical Classifications

Neumann[22] described a classification for computer risk factors (see Table 1).

We find this list too broad in some respects and too narrow in others to be useful for our purposes. Hardware factors are outside the scope of this paper

[12] Roman Yampolskiy, 'Leakproofing the Singularity: The Artificial Intelligence Confinement Problem' *(2012) 19* Journal of Consciousness Studies 1-2

[13] Eliezer Yudkowsky, 'Retrieved from The AI-Box Experiment' (2002) < http://yudkowsky.net/singularity/aibox> accessed 20 January 2020

[14] (n 5)

[15] Daniel Crevier, *AI: The Tumultuous Search for Artificial Intelligence (*Basic Books 1993)

[16] CIO Dive, 'Gartner Serves up 2018 Hype Cycle with a Heavy Side of AI' <https://www.ciodive.com/news/gartner-serves-up-2018-hype-cycle-with-a-heavy-side-of-ai/530385/> accessed 1 November 2019

[17] Nancy Leveson, *Safeware: System Safety and Computers (*Addison-Wesley 1995)

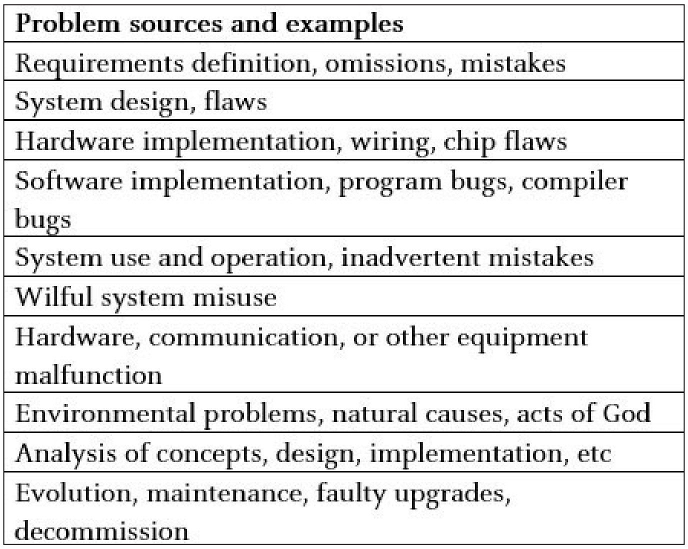
[18] Dan Craigen, Nadia Diakun-Thibault, and Randy Purse, 'Defining Cybersecurity' *(2012)* Technology Innovation Management Review 13-21

[19] Roman Yampolskiy, 'Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures' (2016) *arXiv:1610.07997v1 [cs.AI]*

[20] Shane Legg and Marcus Hutter, 'A Collection of Definitions of Intelligence' (2007) *IDSIA-0707 Technical Report*

[21] Bertrand Meyer, 'John McCarthy' (*Communications of the ACM,* 28 October 2011) <https://cacm.acm.org/blogs/blog-cacm/138907-john-mccarthy/fulltext> accessed 20 January 2020

[22] Peter Neumann, *Computer-Related Risks (*Addison-Wesley 1994)



Problem sources and examples
Requirements definition, omissions, mistakes
System design, flaws
Hardware implementation, wiring, chip flaws
Software implementation, program bugs, compiler bugs
System use and operation, inadvertent mistakes
Wilful system misuse
Hardware, communication, or other equipment malfunction
Environmental problems, natural causes, acts of God
Analysis of concepts, design, implementation, etc
Evolution, maintenance, faulty upgrades, decommission

Table 1: Computer Risk Factors Sources and Examples. Source: P.G. Neumann

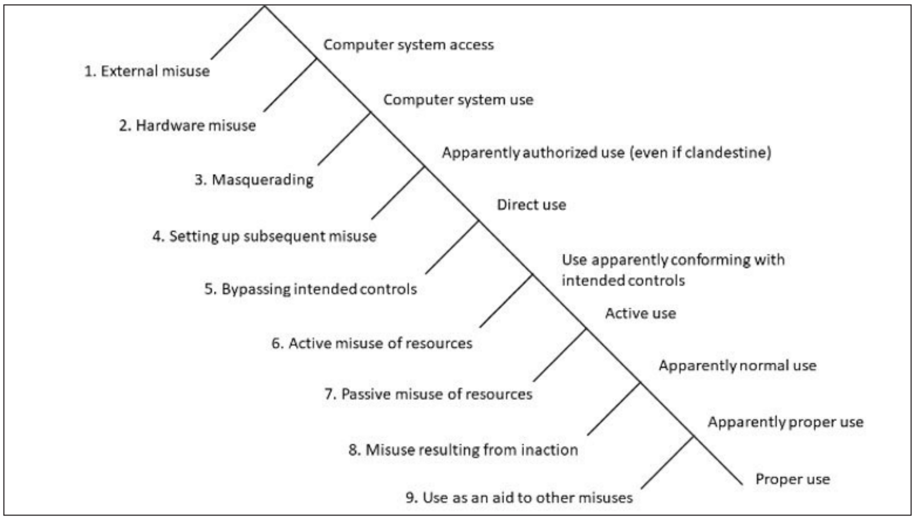

Computer system access
1. External misuse
Computer system use
2. Hardware misuse
Apparently authorized use (even if clandestine)
3. Masquerading
Direct use
4. Setting up subsequent misuse
Use apparently conforming with intended controls
5. Bypassing intended controls
Active use
6. Active misuse of resources
Apparently normal use
7. Passive misuse of resources
Apparently proper use
8. Misuse resulting from inaction
Proper use
9. Use as an aid to other misuses

Figure 1: Classes of Computer Misuse
Source: Neumann and Parker
Note: The leftward branches all involve misuse; the rightward branches represent potentially acceptable use – until a leftward branch is taken.

(and are increasingly irrelevant as software becomes more platform-independent and mobile); software factors need greater elaboration. Neumann and Parker[23] listed classes of computer misuse techniques (see Figure 1).

Despite the tree structure, this represents a system of descriptors rather than a taxonomy in that a given misuse may involve multiple techniques within several classes. The leftward branches all involve misuse; the rightward branches represent potentially acceptable use–until a leftward branch is taken. However, the term 'misuse' implies deliberate agency and

[23] Peter Neumann and Donald Parker, 'A Summary of Computer Misuse Techniques' (1989) *12th National Computer Security Conference* 396-407

Table 2: AI Failure Consequences at Human Aggregation Levels – Schema Tags

Human Aggregation Scale	Consequences					
	Physical	Mental	Emotional	Financial	Social	Cultural
Individual	CIP	CIM	CIE	CIF		
Corporation				CCF		CCC
Community				CYF	CYS	CYC

Table 3: AI Failure Levels of Agency – Schema Tags

Agency	Code
Accidental	AA
Negligent	AN
Innocuous	AI
Malicious	AM

thereby ignores a multitude of failure modes that stem from accidental oversights.

2. AI Failure Classification Dimensions

Here we modify and extend earlier work by Yampolskiy[24] in classifying AI risk factors. Hollnagel[25] deconstructs safety in the steps of *phenomenology* (observables), *etiology* (causes), and *ontology* (nature). We address each of these steps in proposing the following dimensions as useful classification criteria for AI failures:

- Consequences (phenomenology)
- Agency (etiology)
- Preventability (ontology)
- Stage of introduction in the product lifecycle (phenomenology and ontology)

Each will be denoted with a 2- or 3-letter code that we will tag our examples with.

a. Consequences

Consequences may be considered on the scale of human aggregation on which they can occur (see Table 2).

Individuals can range in number from one to every member of the human race; the grouping will be used to denote at what type of aggregation the action of the failure was aimed rather than the number of instances affected. Corporations are legal structures for doing business, of any size. Communities are groupings of people organised for purposes other than business and range from families to nations.

- Physical consequences occur to individuals and may range from inconvenience to loss of life.
- *Mental* consequences occur to individuals and include the alteration of mental states such as beliefs, with concomitant changes in behaviour. For instance, the purpose or effect of 'fake news' is to cause such changes.[26]
- Emotional consequences occur to individuals and include depressive states resulting from AI incidents with physical or mental consequences, and AI usurping roles that people have assumed to be unassailable.
- Financial consequences occur to individuals, corporations, and communities.
- Social consequences are the modifications of behaviour of systems or organisations of people.
- *Cultural* consequences are the modifications of an organisation or grouping's vision, values, norms, systems, symbols, language, assumptions, beliefs, and habits.[27]

Consequences are not necessarily negative, or may be negative in some respects while being positive in others. A superintelligence that enslaved humans in boot camps might keep them in optimal physical condition but pessimal emotional state.

[24] Roman Yampolskiy, 'Taxonomy of Pathways to Dangerous AI' *Proceedings of 2nd International Workshop on AI, Ethics and Society (AIEthicsSociety 2016)* 143-148

[25] Erik Hollnagel, *Safety-I and Safety-II* (Ashgate Publishing 2014)

[26] David M. J. Lazer et al, 'The Science of Fake News' (2018) 359 Science 1094-1096

[27] David Needle, *Business in Context: An Introduction to Business and Its Environment (*Cengage Learning EMEA 2014)

Table 4: AI Failure Degree of Preventability – Schema Tags

Degree of Preventability	Code
Trivially preventable	PT
Preventable with some difficulty	PS
Preventable with great difficulty	PD
Unpreventable	PU

b. Agency

The *agency* of a failure is the degree of human intentionality in its origin or propagation (see Table 3).

An *accidental* failure is one that was not foreseen and could not reasonably have been foreseen. We are departing slightly from the customary engineering definition of 'accident' here in order to draw a more useful distinction. Leveson[28] defines 'accident' as 'An undesired and unplanned (but not necessarily unexpected) event that results in (at least) a specified level of loss.' Thus automobile accidents are foreseeable but neither expected nor desired. We prefer instead to define a *negligent* failure as one that was not foreseen but could (and perhaps should) have been foreseen.

An *innocuous* failure is one deliberately caused, but not with malicious intent, possibly with the intent of causing a more benign effect than what actually resulted. A *malicious* failure is one that was initiated with the intention of causing deleterious effects, whether they were specifically the effects that actually resulted or others. No connection with legal definitions of these terms should be inferred from their attribution to specific events.

c. Preventability

Levels of agency are independent of the *degree of preventability* (see Table 4).

Some failure modes of superintelligences are forecast by some authorities to be unpreventable: '[W]e have seen enough to conclude that scenarios in which some machine intelligence gets a decisive strategic advantage are to be viewed with grave concern.'[29]

Table 5: Software Development Lifecycle Stages – Schema Tags

Lifecycle Stage	Code
Concept	LC
Design	LD
Development	LE
Testing	LT
Operation	LO
Decommissioning	LG

d. Lifecycle Stage

A common taxonomy for computer system errors is the software development *lifecycle stage* (see Table 5); it is often asserted that the cost of fixing an error at each stage is ten times the cost of fixing it in the previous stage.[30]

We add in the less commonly included stages of concept (was it a good idea to do this in the first place?) at the beginning, and decommissioning (what are the problems caused by getting rid of the product) at the end. A superintelligence might be highly resistant to decommissioning.[31]

IV. AI Failures

With these dimensions in mind we now examine various reported and hypothesised failures. Note that there is an unavoidable degree of subjective variability in the classifications of preventability and agency.

[28] (n 17)

[29] (n 5) 154

[30] Maurice Dawson, Darrell Norman Burrell and Emad Rahim, 'Integrating Software Assurance into the Software Development Life Cycle (SDLC)' (2010) 3 Journal of Information Systems Technology and Planning 49-53

[31] '2001: A Space Odyssey (1968) - I'm Sorry, Dave Scene' *(YouTube)* <https://www.youtube.com/watch?v=Wy4EfdnMZ5g> accessed on 1 November 2019

1. Reported Failures

Whereas Yampolskiy[32] enumerated several dozen failures in a timeline that highlighted an exponentially increasing frequency and severity, nearly all of the examples we cite here occurred within the 2016-2019 period and so a chronological ordering would not be illuminating. We will therefore place them instead within a more narrative structure.

The most recognisable and straightforward class of failures result in physical injury to humans, going back to the classic Therac-25 radiation therapy overdose cases[33] (CIP, AN, PS, LD, LE, LO). When an Amazon warehouse robot accidentally punctured a container of bear spray[34] (CIP, AN, PS, LT) it was a more benign outcome of an industrial accident than when a Chinese factory worker was impaled with ten foot-long spikes[35] (CIP, AN, PT, LD). But these and other more fatal accidents with industrial robots going back at least to 1984 when an operator was killed by a 2,500 lb robot that came behind him with no warning[36] (CIP, AN, PS, LD) indicate lack of consideration for humans sharing the same location as machines. A car production plant robot grabbed a worker instead of a part and crushed him against a metal plate, killing him[37] (CIP, AN, PS, LD).

Incidents of cars in semi-autonomous operation causing fatalities include an Uber incorrectly classifying a pedestrian as a false positive match because too many reactions to actual false positives resulted in a jerky ride[38] (CIP, AN, PS, LT), and a Tesla crashing after requesting driver intervention[39] (CIP, CCF, AA, PD, LE).

In medicine, IBM's Watson recommended 'unsafe' cancer treatments[40] (CIP, CCF, AA, PD, LT, LO), and a study of 14 years of robotic surgery concluded that 'a non-negligible number of [preventable] technical difficulties and complications are still being experienced during procedures.'[41] (CIP, CCF, AA, PS, LD).

AI accidents may result in direct financial loss. The May 2010 'Flash Crash' resulted in the Dow Jones Industrial Average dropping about 9% for 36 minutes and resulted from program trading algorithms being inadequately prepared to deal with large volumes of strategically-placed trades which themselves were computer-mediated malice[42] (CIF, CCF, AA, AM, PD, LD, LT). Remediation efforts did not prevent more flash crashes in 2015.[43]

A major concern in the application of AI is privacy. Consumer devices connected to corporate clouds of identity data come under scrutiny, especially when, for instance, an Amazon Alexa node recorded a private conversation and sent it to a random contact[44] (CIE, AA, PS, LT, LO), or an iPhone bug allowed users to listen on others' conversations via FaceTime[45] (CIE, AA, PS, LT). In some cases, the technol-

[32] Roman Yampolskiy, 'Predicting Future AI Failures from Historic Examples' (2018) *foresight* 138-152

[33] J.A. Rawlinson, 'Report on the Therac-25' OCTRT/OCI Physicists Meeting (Kingston, Ontario 1987)

[34] Saqib Shah, 'Amazon Workers Hospitalized after Warehouse Robot Releases Bear Repellent' (*engadget*, 6 December 2018) <https://www.engadget.com/2018/12/06/amazon-workers-hospitalized-robot> accessed 20 January 2020

[35] Tariq Tahir, 'Factory Robot Impales Worker with 10 Foot-long Steel Spikes after Horror Malfunction' (*The Sun,* 14 December 2018) <https://www.thesun.co.uk/news/7954270/factory-robot-malfunctions-and-impales-worker-with-10-foot-long-steel-spikes/> accessed 1 November 2019

[36] John G. Fuller, 'Death by Robot' (1984) Omni, 45-46, 97-102

[37] Associated Press in Berlin, 'Robot Kills Worker at Volkswagen Plant in Germany' (*The Guardian*, 2 July 2015) <https://www.theguardian.com/world/2015/jul/02/robot-kills-worker-at-volkswagen-plant-in-germany> accessed 1 November 2019

[38] Timothy B. Lee, 'Software Bug Led to Death in Uber's Self-driving Crash' (*ars Technica*, 7 May 2018) <https://arstechnica.com/tech-policy/2018/05/report-software-bug-led-to-death-in-ubers-self-driving-crash/> accessed 20 January 2020

[39] Faiz Siddiqui, 'NTSB "Unhappy" with Tesla Release of Investigative Information in Fatal Crash' (*Washington Post*, 1 April 2018) <https://www.washingtonpost.com/news/dr-gridlock/wp/2018/04/01/ntsb-unhappy-with-tesla-release-of-investigative-information-in-fatal-crash/> accessed 1 November 2019

[40] Casey Ross and Ike Swetlitz, 'IBM's Watson Supercomputer Recommended "Unsafe and Incorrect" Cancer Treatments, Internal Documents Show' (*Stat News*, 25 July 2018) <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/> accessed 1 November 2019.

[41] Homa Alemzadeh, Jaishankar Raman, Nancy Leveson, Zbigniew Kalbarczyk and Ravishankar K. Iyer, 'Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data' (2016) 11 *PLoS ONE*

[42] '2010 Flash Crash' (*Wikipedia*) <https://en.wikipedia.org/wiki/2010_Flash_Crash> accessed 1 November 2019

[43] Cory Mitchell, 'The Two Biggest Flash Crashes of 2015' (*Investopedia*, 25 June 2019) <https://www.investopedia.com/articles/investing/011116/two-biggest-flash-crashes-2015.asp> accessed 1 November 2019.

[44] Gary Horcher, 'Amazon Alexa Recorded Private Conversation, Sent it to Random Contact, Woman Says' (*WHBQ*, 24 May 2018) <https://www.fox13memphis.com/news/trending-now/amazon-alexa-recorded-private-conversation-sent-it-to-random-contact-woman-says/755720160> accessed 1 November 2019

[45] Mark Gurman, 'Apple Bug Lets iPhone Users Listen in on Others Via FaceTime' (*Bloomberg*, 28 January 2019) <https://www.bloomberg.com/news/articles/2019-01-29/apple-bug-lets-iphone-users-listen-in-on-others-via-facetime> accessed 1 November 2019

ogy facilitated a casual violation of privacy such as when Uber users' locations and identities were displayed on a screen at a launch party[46] (CIE, AI, PT, LC).

Privacy violations carry more serious consequences when they become misidentifications. The ACLU demonstrated that when they showed that Amazon facial recognition would flag certain members of Congress as wanted criminals[47] (CYF, CYS, AN, PS, LD). A lack of training data (and implicit bias) resulted in facial recognition systems being unable to see black people[48] or tagging them as gorillas[49] (CIE, CYS, AN, PD, LD). Facial recognition used by police in the United Kingdom has been recorded making many false positive identifications[50] (CIE, CIF, CYF, CYS, AN, PS, LT, LO). And in China, facial recognition systems deployed for automated misdemeanor ticketing publicly shamed a woman as a jaywalker when mistaking her photo on the side of a bus for the woman herself[51] (CIE, CIF, AN, PD, LE) and a driver was ticketed for using a cellphone when he was actually scratching his face[52] (CIE, CIF, AN, PD, LE). Traffic cameras in New Orleans ticketed parked cars for speeding[53] (CIF, AN, PS, LD, LO). A man was falsely arrested after systems at Apple misidentified him as stealing from its stores[54] (CIP, CIM, CIE, CIF, CCF, AA, PS, LE).

Not all misidentifications result in such obvious harm. Artist Tom White specialises in creating abstract (and very unarousing) art that is flagged as unacceptable nudity by social media AI.[55] This machine myopia indicates that the development of useful image censorship is not yet realised and some inoffensive art is suppressed. (CIM, CIF, AA, PD, LD).

Implicit misidentification by category is *bias,* another topic of great concern in AI development. With good reason: a report concluded that AIs trained on hiring decisions would replicate or amplify human bias,[56] Amazon's hiring AI turned out to be sexist,[57] and the COMPAS system used in Wisconsin to predict recidivism was biased against blacks[58] (CIE, CIF, CYC, AA, PD, LE). Just as human bias often results from inadequate exposure to diversity, AI bias often arises from the same cause. An attempt to use AI to objectively judge an online international beauty contest without human bias failed when only one of 44 winners it chose had dark skin, prompting speculation that this was due to the training database having few dark faces[59]. And the New Zealand automated passport application checking system rejected an

[46] Kashmir Hill, '"God View": Uber Allegedly Stalked Users For Party-Goers' Viewing Pleasure' (*Forbes*, 3 October, 2014) <https://www.forbes.com/sites/kashmirhill/2014/10/03/god-view-uber-allegedly-stalked-users-for-party-goers-viewing-pleasure/#4b7dd5593141> accessed 1 November 2019

[47] Cyrus Farivar, 'Amazon's Recognition Messes Up, Matches 28 Lawmakers to Mugshots' (*ars Technica*, 26 July 2018) <https://arstechnica.com/tech-policy/2018/07/amazons-rekognition-messes-up-matches-28-lawmakers-to-mugshots/> accessed 1 November 2019

[48] 'Algorithmic Justice League' <https://www.ajlunited.org/> accessed 1 November 2019

[49] Jana Kasperkevic, 'Google Says Sorry for Racist Auto-tag in Photo App' (*The Guardian*,1 July 2015) <https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app> accessed 1 November 2019

[50] Matt Burgess, 'Facial Recognition Tech Used by UK Police is Making a Ton of Mistakes' (*Wired UK*, 4 May 2018) <https://www.wired.co.uk/article/face-recognition-police-uk-south-wales-met-notting-hill-carnival> accessed 1 November 2019

[51] Xinmei Shen, 'Facial Recognition Camera Catches Top Vusinesswoman "Jaywalking" Because her Face Was on a Bus' (*Abacus News*, 22 November 2018) <https://www.abacusnews.com/digital-life/facial-recognition-camera-catches-top-businesswoman-jaywalking-because-her-face-was-bus/article/2174508> accessed 1 November 2019

[52] WTF, 'Chinese Driver Fined for Scratching his Face after Passing AI Traffic Camera' (*9gag*, 26 May 2019) <https://9gag.com/gag/av8VBdd/chinese-driver-fined-for-scratching-his-face-after-passing-ai-traffic-camera> accessed 1 November 2019

[53] Willie James Inman, 'Traffic Camera in New Orleans Giving Speeding Tickets to Parked Cars' (*Fox News*, 11 April 2018) <https://www.foxnews.com/auto/traffic-camera-in-new-orleans-giving-speeding-tickets-to-parked-cars> accessed 1 November 2019

[54] 'Apple AI Accused of Leading to Man's Wrongful Arrest' (*BBC News*, 23 April 2019) <https://www.bbc.com/news/technology-48022890> accessed 1 November 2019

[55] Jason Bailey, 'AI Artists Expose "Kinks" In Algorithmic Censorship' (*Artnome*, 11 December 2018) <https://www.artnome.com/news/2018/12/6/ai-artists-expose-kinks-in-algorithmic-censorship> accessed 1 November 2019

[56] Nicol Turner Lee, Paul Resnick et al, 'Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms' (*Brookings Institute*, 22 May 2019) <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/> accessed 1 November 2019

[57] James Cook, 'Amazon Scraps "Sexist AI" Recruiting Tool that Showed Bias Against Women' (*The Telegraph*, 10 October 2018) <https://www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/> accessed 1 November 2019

[58] Ed Yong, 'A Popular Algorithm Is No Better at Predicting Crimes Than Random People' (*The Atlantic*, 17 January 2018) <https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/> accessed 1 November 2019

[59] Jordan Pearson, 'Why An AI-Judged Beauty Contest Picked Nearly All White Winners' (*Vice*, 5 September 2016) <https://www.vice.com/en_us/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners> accessed 1 November 2019

Asian applicant's photograph, claiming that 'Subject's eyes are closed.'[60] A study demonstrated that implicit race and gender biases in training corpora flowed through into AIs trained on those corpora.[61]

In the hands of an authoritarian regime, AI can create environments prompting comparisons with Orwell's *1984*. Nowhere is this more apparent than in China, which has embraced facial recognition on a large scale.[62] AI there blocks mention of the Tiananmen Square massacre on social media[63] (CYS, AI, PT, LC). While this software is being used to create exactly its intended effect, we label this a failure because it has consequences many western observers would consider to be socially harmful. China has a 'social credit' scoring system reminiscent of a Black Mirror episode,[64] linked to social media and consumer systems such as Sesame Credit,[65] that will ban people from certain venues like flights and hotels for poor scores, which may be incurred by undesirable behaviour such as buying video games (CYC, AI, PT, LC). Some commentators speculate that this will have consequences in health care.[66] Also in China, AI is being used to grade school papers,[67] with some good writing being given poor marks (CYS, AI, PD, LO). And AI is used to monitor the moods of workers[68] and the attention paid by children in class[69] with the most attentive being rewarded (CIM, CYC, AI, PT, LC).

In the West the dangers are more nascent. Researchers at the University of Pennsylvania demonstrated that textual analysis of an individual's Facebook posts could predict 21 different medical conditions such as diabetes.[70] Others showed that AI was better than people at determining sexual orientation from a photograph,[71] while a third group determined that AI could detect certain genetic diseases from faces.[72] A Department of Homeland Security program predicts which flyers are potential terrorists[73] from demographic and travel data alone, and if those travellers make it to the European Union they may face an AI-powered lie detection system at the border.[74] The startup Faception claims its software can predict personality traits such as pedophile or poker player from facial image analysis, causing one commentator to liken it to phrenology.[75] A person's gait

[60] James Titcomb, 'Robot Passport Checker Rejects Asian Man's Photo for Having his Eyes Closed' (*The Telegraph, 7 December 2016*) <https://www.telegraph.co.uk/technology/2016/12/07/robot-passport-checker-rejects-asian-mans-photo-having-eyes/> accessed 1 November 2019

[61] Aylin B. Caliskan, 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases' (2017) Science 183-186

[62] Sijia Jiang, 'Backing Big Brother: Chinese Facial Recognition Firms Appeal to Funds' (*Reuters*, 12 November 2017) <https://www.reuters.com/article/us-china-facialrecognition-analysis/backing-big-brother-chinese-facial-recognition-firms-appeal-to-funds-idUSKBN1DD00A> accessed 1 November 2019

[63] Michael Grothaus, 'Now AI Easily Erases the Tiananmen Square Massacre from Online Memory' (*Fast Company*, 28 May 2019) <https://www.fastcompany.com/90355806/now-ai-easily-erases-the-tiananmen-square-massacre-from-online-memory> accessed 1 November 2019

[64] J. Wright (Director) (2016) *Black Mirror: Nosedive* [Motion Picture]

[65] 'Social Credit System' (*Wikipedia*) <https://en.wikipedia.org/wiki/Social_Credit_System> accessed 1 November 2019

[66] John Harris, 'The Tyranny of Algorithms is Part of our Lives: Soon They Could Rate Everything We Do ' (*The Guardian*, 5 March 2018) <https://www.theguardian.com/commentisfree/2018/mar/05/algorithms-rate-credit-scores-finances-data> accessed 1 November 2019

[67] Stephen Chen, 'China's Schools are Quietly Using AI to Mark Students' Essays ... But do the Robots Make the Grade?' (*South China Morning Post*, 27 May 2018) <https://www.scmp.com/news/china/society/article/2147833/chinas-schools-are-quietly-using-ai-mark-students-essays-do> accessed 1 November 2019

[68] Jamie Fullerton, '"Mind-reading" Tech Being Used to Monitor Chinese Workers' Emotions' (*The Telegraph*, 30 April 2018) <https://www.telegraph.co.uk/news/2018/04/30/mind-reading-tech-used-monitor-chinese-workers-emotions/> accessed 1 November 2019

[69] Jiayun Feng, 'Chinese Parents Want Students to Wear Dystopian Brainwave-detecting Headbands' (*supChina*, 5 April 2019) <https://supchina.com/2019/04/05/chinese-parents-want-students-to-wear-dystopian-brainwave-detecting-headbands/> accessed 1 November 2019

[70] Raina M. Merchant et, 'Evaluating the Predictability of Medical Conditions from Social Media Posts' (2019) PLoS ONE

[71] Yilun Wang and Michal Kosinski, 'Deep Neural Networks are More Accurate than Humans at Detecting Sexual Orientation from Facial Images (2018) Journal of Personality and Social Psychology 246–257

[72] Nina Avramova, 'AI Technology Can Identify Genetic Diseases by Looking at Your Face, Study Says' (*CNN*, 8 January 2019) <https://edition.cnn.com/2019/01/08/health/ai-technology-to-identify-genetic-disorder-from-facial-image-intl/index.html> accessed 1 November 2019

[73] Sam Biddle, 'Homeland Security Will Let Computers Predict Who Might Be a Terrorist on Your Plane — Just Don't Ask How It Works' (*The Intercept*, 3 December 2018) <https://theintercept.com/2018/12/03/air-travel-surveillance-homeland-security/> accessed 1 November 2019

[74] 'Smart Lie-detection System to Tighten EU's Busy Borders' (*European Commission*, 24 October 2018) <https://ec.europa.eu/research/infocentre/article_en.cfm?artid=49726> accessed 1 November 2019

[75] Matt McFarland, 'Terrorist or Pedophile? This Start-up Says it Can Out Secrets by Analyzing Faces' (*Washington Post,* 24 May 2016) <https://www.washingtonpost.com/news/innovations/wp/2016/05/24/terrorist-or-pedophile-this-start-up-says-it-can-out-secrets-by-analyzing-faces/> accessed 1 November 2019

can be used to identify them.[76] Two systems supply 'predictive policing' systems that, inviting a comparison with the movie *Minority Report,* suggest where crime is likely to occur.[77] [78] Companies exploit human psychology to get our attention,[79] the US military studies how to influence Twitter users,[80] and the Pentagon wants to predict protests against the US President via social media surveillance.[81]

As Yampolskiy[82] pointed out, 'An AI designed to do X will eventually fail to do X,' codified as the *Fundamental Theorem of Security:* There is no such thing as a 100% secure system. In all the examples in the previous paragraph the latent failures are the ones implied by this theorem, with their concomitant risks.

The consequences of *misinformation* spread by AI include of course 'fake news,' such as that attributed to Cambridge Analytica,[83] assiduously spread by social media,[84] and 'deep fake' videos,[85] which could be used to automate blackmail at scale[86] (CIM, CYS, AM, PU, LO)

A class of incidents illustrates that much AI is not yet mature. A hardware design bug allowed memory protection violations in years' worth of Intel chips[87] (CCF, AA, PD, LD). And Microsoft's Tay chatbot became racist within hours of being deployed to learn from other Twitter users[88] (CYS, AA, PS, LC/LD). Some aspects of this immaturity are fundamentally brittle; for instance, when a digital exchange lost $137 million because the one person holding the master password died,[89] or when bots tasked with maintaining Wikipedia fought with each other for years,[90] Deep reinforcement learning fails more often than admitted.[91]

Intentional misuse spans many incidents; to cite two, smart scooters for hire were hacked to display obscene messages and be used without payment[92] (CCF, AM, PS, LE) and Domino's Pizza affiliation app was fooled into granting points by fake pictures of pizza[93] (CCF, AM, PD, LC).

Some 'backfire' events result in damage to the AI industry through overreaching or misrepresentation. For instance, a preternaturally capable healthcare AI called 'Zach' in New Zealand was suspected to be a person in disguise.[94] And the Sophia robot attracts a

[76] 'Forensic Gait Analysis' (*Royal Society of Edinburgh*, November 2017) <https://royalsociety.org/~/media/about-us/programmes/science-and-law/royal-society-forensic-gait-analysis-primer-for-courts.pdf> accessed 1 November 2019

[77] 'PredPol' <https://www.predpol.com/> accessed 1 November 2019

[78] 'Palantir' <http://www.palantir.com/> accessed 1 November 2019

[79] Tristan Harris, 'How a Handful of Tech Companies Control Billions of Minds Every Day' (*TED*, April 2017) <https://www.ted.com/talks/tristan_harris_the_manipulative_tricks_tech_companies_use_to_capture_your_attention> accessed 1 November 2019

[80] Ben Quinn and James Ball, 'US Military Studied How to Influence Twitter Users in Darpa-funded Research' (*The Guardian*, 8 July 2014) <https://www.theguardian.com/world/2014/jul/08/darpa-social-networks-research-twitter-influence-studies> accessed 1 November 2019

[81] Nafeez Ahmed, 'Pentagon Wants to Predict Anti-Trump Protests Using Social Media Surveillance' (*Vice*, 30 October 2018) <https://www.vice.com/en_us/article/7x3g4x/pentagon-wants-to-predict-anti-trump-protests-using-social-media-surveillance> accessed 1 November 2019

[82] (n 19)

[83] 'Cambridge Analytica planted fake news' (*BBC*, 20 March 2018) <https://www.bbc.com/news/av/world-43472347/cambridge-analytica-planted-fake-news> accessed 1 November 2019

[84] Emerging Technology from the arXiv, 'First Evidence That Social Bots Play a Major Role in Spreading Fake News' (*Technology Review*, 7 August 2017) <https://www.technologyreview.com/s/608561/first-evidence-that-social-bots-play-a-major-role-in-spreading-fake-news/> accessed 1 November 2019

[85] Ben Collins, 'This Viral Schwarzenegger Deepfake isn't Just Entertaining. It's a Warning' (*NBC News*, 12 June, 2019) <https://www.nbcnews.com/tech/tech-news/viral-schwarzenegger-deepfake-isn-t-just-entertaining-it-s-warning-n1016851> accessed 1 November 2019

[86] Paul Bricman, 'DeepFake Ransomware' (*Medium*, 2 February 2019) <https://medium.com/@paubric/deepfake-ransomware-oaas-part-1-b6d98c305cd9> accessed 1 November 2019

[87] Zack Whittaker, 'New Secret-spilling Flaw Affects Almost Every Intel Chip Since 2011' (*Tech Crunch*, 14 May 2019) <https://techcrunch.com/2019/05/14/zombieload-flaw-intel-processors/> accessed 1 November 2019

[88] Elle Hunt, 'Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter' (*The Guardian*, 24 March 2016) <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter> accessed 1 November 2019

[89] Dan Goodin, 'Digital Exchange Loses $137 Million as Founder Takes Passwords to the Grave' (*ars Technica*, 2 February 2019) <https://arstechnica.com/information-technology/2019/02/digital-exchange-loses-137-million-as-founder-takes-passwords-to-the-grave/> accessed 1 November 2019

[90] Sara Chodosh, 'Wikipedia Bots Spent Years Fighting Silent, Tiny Battles with Each Other' (*Popular Science*, 27 February 2017) <https://www.popsci.com/wikipedia-bots-fighting/> accessed 1 November 2019

[91] Alex Irpan, 'Deep Reinforcement Learning Doesn't Work Yet' (*Sorta Insightful*, 14 February 2018) <https://www.alexirpan.com/2018/02/14/rl-hard.html> accessed 1 November 2019

[92] Matt Novak, 'Lime Scooters Hacked to Say Sexual Things to Riders in Australia' (*Gizmodo*, 24 April 2019) <https://gizmodo.com/lime-scooters-hacked-to-say-sexual-things-to-riders-in-1834264534> accessed 1 November 2019

[93] Matthew Gault, 'Take Pictures of Fake Pizzas to Get a Free Pizza from Domino's' (*Vice*, 6 March 2019) <https://www.vice.com/en_us/article/kzdkgw/take-pictures-of-fake-pizzas-to-get-a-free-pizza-from-dominos> accessed 1 November 2019

[94] David Farrier, 'The Mystery of Zach, New Zealand's all-too-miraculous medical AI' (*The Spinoff*, 6 March 2018) <https://thespinoff.co.nz/the-best-of/06-03-2018/the-mystery-of-zach-new-zealands-all-too-miraculous-medical-ai/> accessed 1 November 2019

degree of adulation far beyond its real capabilities.[95] The threat here is to the reputation of AI and its community (CYF, CYC, AI, PU, LO).

AI that is unintentionally insensitive also damages its own reputation, such as the AI that thought that house burning down was 'spectacular'[96] and Paypal's virtual assistant which insensitively replied, 'Great!' when someone told it, 'I got scammed'[97] (CIE, AA, PS, LE). The Starbucks shift-scheduling software was also insensitive when it optimised for hour-by-hour business needs but assigned workers to unpredictable and erratic schedules[98] (CIP, CIF, CCC, AN, PT, LD). AI that is trusted without verification may not live up to that trust, such as when a model used to grade the 'value-add' imparted by New York City teachers was found to generate essentially random results[99] (CIM, CIE, CIF, AN, PS, LT). A corporate employment workflow system was unstoppable in terminating an employee erroneously flagged as superfluous;[100] after three weeks spent fixing the error he declined to return to the firm (CIE, CIF, CCF, AA, PS, LD).

Some failures are so benign on the surface that many casual observers would classify them as cute behaviour rather than failures. When a robot (with smiley face to boot) on the International Space Station stopped obeying astronauts[101] the parallels with HAL 9000 of *2001: A Space Odyssey* were so irresistible as to obscure the real risks of a computer failure in a critical environment. Apple's Siri's initial response to the request 'Call me an ambulance' was to refer to the user thereafter as 'ambulance'[102] (CIP, AA, PS, LE). When a text generator created weird descriptions of Bitcoin[103], and an AI's predicted YouTube pornography searches,[104] the results were so funny as to be equally disarming (CYC, AA, PS, LD). A trivial typo in the code for a game agent made it much easier to beat than it should have been[105] (CIM, AA, PT, LT). A Roomba spread dog poop all over a house[106] (CIP, CIF, AA, PS, LE). A sign printed in Welsh translated to 'I am not in the office at the moment. Send any work to be translated.'[107] (CYC, AN, PT, LO). The 'swarm intelligence' UNU failed to predict the results of the Kentucky Derby the second time around after previously winning the superfecta.[108] (CIF, AA, PD, LE). A neural network hallucinated sheep in images where there were none, or mislabelled them when they were placed in (admittedly unusual) locations[109] (CIM, AA, PS, LE). And in a story guaranteed to get more laughs than

[95] Noel Sharkey, 'Mama Mia It's Sophia: A Show Robot Or Dangerous Platform To Mislead?' (*Forbes*, 17 November 2018) <https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#4fabeea87ac9> accessed 1 November 2019

[96] Margaret Mitchell, 'How We Can Build AI to Help Humans, Not Hurt Us' (*TED*, October 2017) <https://www.ted.com/talks/margaret_mitchell_how_we_can_build_ai_to_help_humans_not_hurt_us/transcript> accessed 1 November 2019

[97] *Facebook*, 20 March 2019 <https://www.facebook.com/photo.php?fbid=10217225240875399> accessed 1 November 2019

[98] Jodi Kantor, 'Working Anything but 9 to 5' (*New York Times*, 13 August 2014) <https://www.nytimes.com/interactive/2014/08/13/us/starbucks-workers-scheduling-hours.html> accessed 1 November 2019

[99] Gary Rubinstein, 'Analyzing Released NYC Value-Added Data Part 2' (*Teach for Us*, 28 February 2012) <http://garyrubinstein.teachforus.org/2012/02/28/analyzing-released-nyc-value-added-data-part-2/> accessed 1 November 2019

[100] Jane Wakefield, 'The Man Who Was Fired by a Machine' (*BBC News*, 21 June 2018) <https://www.bbc.com/news/technology-44561838> accessed 1 November 2019

[101] Jamie Seidel, 'CIMON, the International Space Station's Artificial Intelligence, Has Turned Belligerent' (*News Corp Australia,* 5 December 2018) <https://www.news.com.au/technology/science/space/cimon-the-international-space-stations-artificial-intelligence-has-turned-belligerent/news-story/953a84bc8c4fe414eed2d0550e1d8bf4> accessed 1 November 2019

[102] Will Knight, 'Tougher Turing Test Exposes Chatbots' Stupidity' (*Technology Review*, 14 July 2016) <https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/> accessed 1 November 2019

[103] Daniel Oberhaus, 'Watch This Hilarious Bitcoin Explainer Generated by an AI' (*Vice*, 23 May 2018) <https://www.vice.com/en_us/article/xwmy9a/watch-botnik-ai-bitcoin-explainer> accessed 01 November 2019

[104] Drew Schwartz, 'AI Predicted the Future of Porn Searches and We Can't Stop Laughing' (*Vice*, 6 March 2018) <https://www.vice.com/en_us/article/bj54xv/ai-predicted-the-future-of-porn-searches-and-we-cant-stop-laughing-vgtrn> accessed 01 November 2019

[105] Sam Machkovech, 'A Years-old, One-letter Typo Led to Aliens: Colonial Marines' Weird AI' (*ars Technica*, 13 July 2018) <https://arstechnica.com/gaming/2018/07/a-years-old-one-letter-typo-led-to-aliens-colonial-marines-awful-ai/> accessed 01 November 2019

[106] Jesse Newton (*Facebook*, 9 August 2016) <https://www.facebook.com/jesse.newton.37/posts/776177951574> accessed 01 November 2019

[107] 'E-mail Error Ends up on Road Sign' (*BBC*, 31 October 2008) <http://news.bbc.co.uk/2/hi/7702913.stm> accessed 01 November 2019

[108] David Z. Morris, 'Artificial Intelligence Fails on Kentucky Derby Predictions' (*Fortune*, 7 May 2017) <https://fortune.com/2017/05/07/artificial-intelligence-kentucky-derby-predictions/> accessed 01 November 2019

[109] Janelle Shane, 'Do Neural Nets Dream of Electric Sheep?' (*AI Weirdness*, 18 March 2018) <https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep> accessed 01 November 2019

fears of AI failure, Alexa devices were alleged to be spontaneously laughing.[110]

Behavior that is also perceived as ‘cute’ in the sense of ‘look at how smart my child is,’ can be more concerning because it indicates just how creative AI can be in solving problems with solutions that eluded humans. AIs ‘cheat’ at games by finding loopholes in the rules or unintended back doors in the implementation.[111] One AI invented (or rediscovered) steganography in order to meet its goals.[112] And GPT2, a text generator developed by OpenAI, an organisation dedicated to open sourcing AI to ensure its safety, was deemed to be so good at what it did that it would be too dangerous to publish its code.[113]

Genetic Algorithms

Genetic algorithms can be so innovative at ‘breaking the rules’[114] [115] that they check every category of failure classification, suggesting a path towards unbounded risk, and are therefore collected in this subsection.

- They can exploit misfeatures or bugs in their environment, such as when in the developmental stages of the NERO video game, players’ robots evolved a wiggling motion that allowed them to walk up walls rather than solve the obstacles ‘properly’ by walking around the walls,[116] or when in a capstone project for a graduate level class, students were required to make a a five-in-a-row Tic-Tac-Toe game played on an infinitely large board. One submission’s algorithm evolved to request non-existent moves that were extremely far away, leading to an automatic win since the other players system would crash.[117]
- They can ‘cheat’ by exploiting loopholes in the rules of their goals, such as in an experiment that involved organisms navigating paths, when one organism created an odometer to allow it to navigate the path precisely and earn a perfect score,[118] or when an attempt to create creatures that could evolve swimming strategies resulted in them learning that by twitching their body parts rapidly, they could obtain more energy that let them swim at unrealistic speeds.[119]
- They can reinvent, to their creators’ surprise, capabilities of biological organisms, such as in an experiment where robotic organisms had to find foods or poisons that were both represented by red lights and could use blue lights to communicate with other robots, the organisms evolved in surprising ways that resembled mimicry and dishonesty in nature,[120] or when a digital evolution model that was initially thought to have been a complete failure, was discovered to have reproduced the biological concept Drake’s rule without having been told to do so.[121]
- They can improvise novel solutions to their assigned tasks, such as when 3-D creatures that could run, walk, and swim were gauged by a fitness function of average ground velocity, which resulted in creatures that were tall and rigid, falling over and

[110] Rachel Sandler, ‘Some Amazon Echo Devices Are Spontaneously Laughing, And Nobody Knows Why’ (*Science Alert*, 7 March 2018) <https://www.sciencealert.com/amazon-echo-devices-are-creepily-laughing-at-people> accessed 01 November 2019

[111] Victoria Krakovna, ‘Specification Gaming Examples in AI’ (2 April 2018) <https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/> accessed 01 November 2019

[112] Devin Coldewey, ‘This Clever AI Hid Data from its Creators to Cheat at its Appointed Task’ (*techcrunch*, 31 December 2018) <https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/> accessed 01 November 2019

[113] Tom Simonite, ‘The AI Text Generator That's Too Dangerous to Make Public’ (*Wired*, 14 February 2019) <https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/> accessed 01 November 2019

[114] Janelle Shane, ‘When Algorithms Surprise Us’ (*AI Weirdness*, 13 April 2018) <https://aiweirdness.com/post/172894792687/when-algorithms-surprise-us> accessed 01 November 2019

[115] Joel Lehman et al, ‘The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities’ (2018) *arXiv:1803.03453v1 [cs.NE]*

[116] Kenneth O. Stanley et al, ‘Real-time Neuroevolution in the NERO Video Game’ (2005) 9 IEEE Transactions on Evolutionary Computation 653–668

[117] David E. Moriarty and Risto Miikkulainen, ‘Forming Neural Networks through Efficient and Adaptive Co-evolution’ (1997) 5 Evolutionary Computation 373–399

[118] Laura M. Grabowski et al, ‘A Case Study of the De Novo Evolution of a Complex Odometric Behavior in Digital Organisms (2013) 8 *PLoS One* e60466

[119] Karl Sims, ‘Evolving Virtual Creatures’ (1994) Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques 15–22

[120] Sara Mitri, ‘The Evolution of Information Suppression in Communicating Robots with Conflicting Interests’ (2009) Proceedings of the National Academy of Sciences 15786–15790

[121] Thomas K. Hindré, ‘New Insights into Bacterial Adaptation Through In Vivo and In Silico Experimental Evolution*’ (2012) 10* Nature Reviews Microbiology 352–365

using their potential energy to gain high velocity,[122] or when a robot arm was programmed to interact with a small box on a table, but the gripper was broken, resulting in the robot hitting the box with the gripper in a way that would force the gripper to hold the box firmly.[123]

- They can creatively exceed their goals, such as when the artificial life system Tierra, not expected to evolve higher life forms for years, created complex ecological systems on the first successful run,[124] or when robots that were designed to detect and travel to a light source evolved a spinning behavior that was more efficient than the expected Braitenberg-style movement.[125]

2. Hypothetical Failures

A video producer depicted a fictional future where an artificial superintelligence charged with copyright enforcement hacked people's brains with nanotechnology to correct violations[126] (CIM, CYC, AI, PD, LO). It was demonstrated that a DNA sequencer could be hacked through (currently non-existent) flaws in a compression algorithm[127] (CIP AM, PD, LE).

Most shows that explore AI failure develop a theme epitomised by *Terminator* series: a massive AI becomes self-aware and attempts to destroy humanity. (CIP, CIE, CIF, CCF, CYF, CYS, CYC, AN, AI, PD, LC, LO). Variations include *Colossus: The Forbin Project,* where the AI imprisons humanity to end conflict (CIM, CIE, CYS, CYC, AN, AI, PD, LC, LO), the same goal as the AI VIKI in the movie *I, Robot* and the robots in Jack Williamson's novelette 'With Folded Hands'.[128] One of the least apocalyptic failures was explored in the 2013 film *Her,* where virtual assistant AIs have unforeseen intimate relationships with many humans who are largely changed for the better (CIE, AA, PD, LD).

Broad Classifications of Future Failure Scenarios

Another classification for failures can be applied to future scenarios.

Figure 2 depicts the severity and scale (number of individuals affected) of broadly classified failure scenarios. In chronological order these are:

1. Autonomous weapons, which currently mostly fall into the 'lethal' category, [129] [130] .
2. Employment automation: The potential segment of the population made jobless through AI automation.
3. Control failures: AI of sufficient complexity and power that bugs cause catastrophes.
4. Conscious AIs: Control failures in AGIs or AIs that are so complex that their behavior is most usefully categorised as 'conscious.'
5. Self-replicating machines: Embodied AIs that can create copies of themselves from raw materials in the environment.

The scenario of 'Conscious AIs' merits some elaboration. Whether an AI is actually conscious is going to become an increasingly difficult and contentious question to answer, but this scenario does not depend on the answer. The 'apparently conscious' AIs in this category are ones that, whether they are conscious or not, will be doing such a good impression of consciousness that it would be more productive to think of them that way than to apply traditional computer science methods to them. We will have reached this stage when the field of 'AI psychiatry' comes into existence.

The chart is not to scale; these are qualitative assessments intended to provoke and inform strategic planning. While some of these labels are apocalyptic, we are motivated by considering Normal Accident Theory[131] and Maas' application to AI: 'At their extreme, unexpected interactions between competing systems, especially in cyberspace, could cause un-

[122] Karl Sims, 'Evolving 3D Morphology and Behavior by Competition' (1994) 1 Artificial Life 353–372

[123] Ecarlat, 'Learning a High Diversity of Object Manipulations Through an Evolutionary-based Babbling' (2015) Proceedings of Learning Objects Affordances Workshop at IROS, 1-2

[124] Thom Ray, 'J'ai Joué à Dieu et Créé la Vie Dans Mon Ordinateur' (1992) Le Temps Stratégique 68–81

[125] Watson, 'Embodied Evolution: Distributing an Evolutionary Algorithm in a Population of Robots' (2002) 39 Robotics and Autonomous Systems 1–18

[126] 'The Artificial Intelligence That Deleted A Century' (*YouTube*) <https://www.youtube.com/watch?v=-JlxuQ7tPgQ> accessed 01 November 2019

[127] Andy Greenberg, 'Biohackers Encoded Malware in a Strand of DNA' (*Wired*, 9 October 2017) <https://www.wired.com/story/malware-dna-hack/> accessed 01 November 2019

[128] Jack Williamson, *With Folded Hands* (Fantasy Press 1947)

[129] 'Slaughterbots' (*YouTube*) <https://www.youtube.com/watch?v=9CO6M2HsoIA> accessed 01 November 2019

[130] 'Ban Lethal Autonomous Weapons' <https://autonomousweapons.org/> accessed 01 November 2019

[131] Charles Perrow, *Normal Accidents: Living with High Risk Technologies (*Princeton University Press 2011)

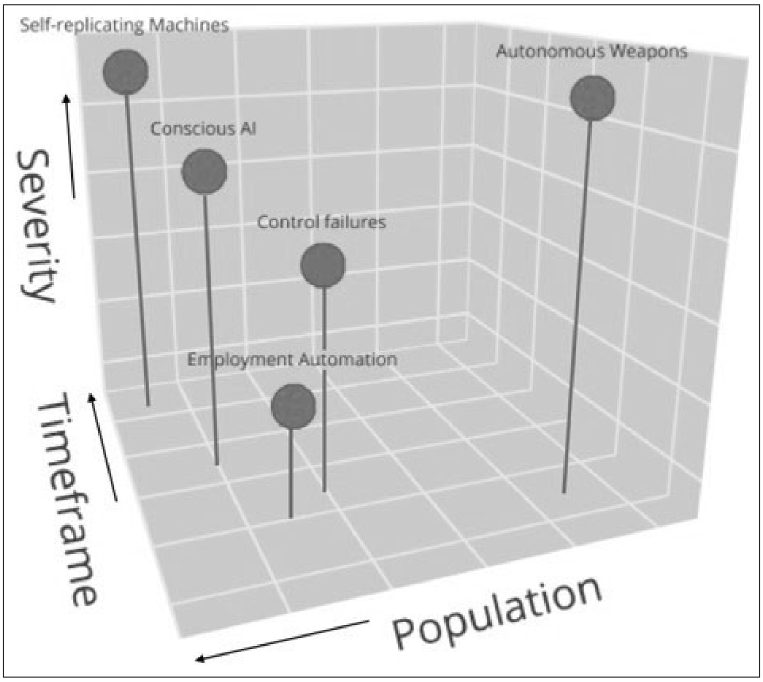


Figure 2: AI failure scenario classes charted by distance into the future, number of humans affected (logarithmic scale), and severity of effect
Source: Authors' elaboration

expected escalation—a 'flash war', analogous to the algorithmic flash crashes observed in the financial sector.'[132]

V. Responses

There are various responses to these failures and risks. Several address privacy. 'Differential privacy' masks individual data in Big Data collections.[133] The Myelin framework preserves privacy in trusted hardware enclaves.[134] Another approach encrypts data before using it to train neural networks without loss of capability.[135] The Data Selfie browser add-on shows leakage of personal data[136]. Another program confuses *ad tracking* by clicking on every ad in the background.[137] A Facebook container isolates your Facebook activity from everything else you do[138] and a program creates search noise to drown out your actual searches.[139]

Defenses are being developed against hacking image recognition networks through microchanges.[140]

[132] Matthijs Maas, Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment (2018) *AIES'8* 223-228

[133] 'Slaughterbots' (*YouTube*) <https://www.youtube.com/watch?v=9CO6M2HsoIA> accessed 01 November 2019

[134] Nick Hynes, 'Efficient Privacy-Preserving ML Using TVM' (*TVM*, 9 October 2018) <https://tvm.ai/2018/10/09/ml-in-tees.html> accessed 1 November 2019.

[135] Morten Dahl, 'Private Image Analysis with MPC Training CNNs on Sensitive Data' (*Cryptography and Machine Learning*, 19 September 2017) <https://mortendahl.github.io/2017/09/19/private-image-analysis-with-mpc/> accessed 1 November 2019

[136] (*Data Selfie*) < https://dataselfie.it/> accessed 1 November 2019

[137] 'AdNauseam Banned from the Google Web Store' (*AdNauseam*, 5 January 2017) <https://adnauseam.io/free-adnauseam.html> accessed 1 November 2019

[138] Dave Camp, 'Firefox Now Available with Enhanced Tracking Protection by Default Plus Updates to Facebook Container, Firefox Monitor and Lockwise' (*The Mozilla Blog*, 4 June 2019) <https://blog.mozilla.org/blog/2019/06/04/firefox-now-available-with-enhanced-tracking-protection-by-default/>> accessed 1 November 2019

[139] Track Me Not < http://trackmenot.io/> accessed 1 November 2019

[140] Ian Goodfellow, Nicolas Papernot et al, 'Attacking Machine Learning with Adversarial Examples' (*OpenAI*, 24 February 2017) <https://openai.com/blog/adversarial-example-research/> accessed 01 November 2019

VI. Conclusions

While we have not made recommendations as to how to address AI failures in each category of the dimensions we have presented, we hope that this classification scheme will make the development of remediation approaches easier.

The importance of this effort may be extrapolated from Leveson's observation that 'The design of the automated system may make the system harder to manage during a crisis.'[141] Noting that this was true of the state of the art in 1995, we are concerned with how systems that are not just far more automated but autonomous may also be far harder to manage during a crisis. The more complex a system becomes, the larger the task X that may be assigned to that system, and so the larger the consequences of the system failing to do X. Today, a humor-generating system writes a joke that isn't funny; tomorrow, employee screening software will hire the wrong people, next week, a system designed to protect a national power grid from cyberattack will fail to do that, etc. Observe that AI systems that perform common human-centric tasks such as image recognition do so in ways that are unrelated to how humans perform those tasks, and are consequentially easily fooled by near-invisible changes;[142] that furthermore AI can operate on completely alien concepts such as the 'opposite' of an image to show, eg, the opposite of a cat.[143] These examples indicate that AI systems used to perform complex human-like tasks will have extremely unpredictable failure modes.

Some people in the AI community view these discussions as scaremongering that impedes the development of AI; to them we quote William Bogard chronicling the Bhopal chemical plant tragedy:

'We are not safe from the risks posed by hazardous technologies, and any choice of technology carries with it possible worst-case scenarios that we must take into account in any implementation decision. The public has the right to know precisely what these worst-case scenarios are and participate in all decisions that directly or indirectly affect their future health and well-being. In many cases, we must accept the fact that the result of employing such criteria may be a decision to forego the implementation of a hazardous technology altogether.'[144]

[141] (n 17)

[142] Avishek Bose and Parham Aarabi, 'Adversarial Attacks on Face Detectors using Neural Net based Constrained Optimization' <https://joeybose.github.io/assets/adversarial-attacks-face.pdf> accessed 1 November 2019

[143] Janelle Shane, 'What is the Opposite of Guacamole? ' (*AI Weirdness*, 10 May 2019) <https://aiweirdness.com/post/184781529122/what-is-the-opposite-of-guacamole> accessed 1 November 2019

[144] William Bogard, *The Bhopal Tragedy (*Westview Press 1989)

An AGI with Time-Inconsistent Preferences

James D. Miller and Roman Yampolskiy*

An artificial general intelligence (AGI) might have time-inconsistent preferences where it knows that it will disagree with the choices its future self will want to make. Such an AGI would not necessarily be irrational. An AGI with such preferences might seek to modify the preferences or constrain the decision making of its future self. Time-inconsistency increases the challenge of building an AGI aligned with humanity's values.

I. Introduction

This paper reveals a trap for artificial general intelligence (AGI) theorists who use economists' standard method of discounting. This trap is implicitly and falsely assuming that a rational AGI would have time-consistent preferences. An agent that realises that it has time-inconsistent preferences knows that its future self will disagree with its current self concerning intertemporal decision making. Such an agent cannot automatically trust its future self to carry out plans that its current self considers optimal.

Economists have long used utility functions to model how rational agents behave.[1] AGI theorists often rely on these utility functions because they assume that an AGI would either start out as rational or modify itself to become rational.[2,3,4,5,6]

When economists model intertemporal decision making, they assume that people place a lower value on receiving money or utility in the future than they do today because people discount future rewards. Economists generally assume that such discounting takes on a particular functional form. Critical for this paper, this functional form causes agents to have time-consistent preferences, and this form does not follow from the assumptions of rationality.

This paper explains why we should model how a future AGI will behave, explores what time-consistent preferences are, discusses why rational AGIs might not have them, and explores how an AGI with time-inconsistent preferences might behave.

II. The Value the Modeling AGIs

Over the next few decades humanity has a good chance of creating computer general intelligences much smarter than us.[7] If these AGIs are friendly towards humanity, they could bring enormous benefit. But a substantial literature claims that the challenge of making these AGIs friendly will range from difficult to near impossible.[8,9,10,11]

An AGI's goals, or what economists would call its utility function, will be determined by its code. We do not yet know how to reduce our values to the language of computer programs. Worse, we might learn how to build powerful AGIs before we learn how to translate our values into the code they run on.

DOI: 10.21552/delphi/2019/4/9

* James D. Miller, for correspondence: <jdmiller@Smith.edu>; Roman V. Yampolskiy, University of Louisville, Kentucky, USA, <roman.yampolskiy@louisville.edu> Supported by Future of Life Grant RFP2-148

[1] Andreu Mas-Colell, Michael Dennis Whinston and Jerry R. Green, Microeconomic Theory (Vol. 1, New York, Oxford University Press 1995)

[2] Stephen M. Omohundro, 'The Basic AI Drives' (2008) 171 AGI 483-492

[3] Eliezer Yudkowsky, 'Complex Value Systems in Friendly AI', International Conference on Artificial General Intelligence (Springer 2011) 388-393

[4] Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (Oxford University Press 2014)

[5] Nate Soares et al, 'Corrigibility' (2015) Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence 2015

[6] Roman V. Yampolskiy, Artificial Superintelligence: A Futuristic Approach (Chapman and Hall/CRC 2015)

[7] Katja Grace et al, 'When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 Journal of Artificial Intelligence Research 729-754

[8] ie Olle Häggström, *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press 2016)

[9] ie (n 6)

[10] ie (n 4)

[11] ie James D. Miller, Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World (BenBella Books Inc. 2012)

For many types of utility function an AGI could have, it would likely have similar instrumental (intermediate) drives.[12] One such drive would be to gather resources.[13] The more resources an AGI had, the more progress it could make towards almost any goal it might have, analogous to how most humans could better achieve their objectives if they had additional wealth. Unfortunately, a powerful AGI might consider the atoms in human bodies to be valuable resources that could be repurposed to fulfilling the AGI's ultimate goals. As AGI theorist Eliezer Yudkowsky has written, 'The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else'[14].

A powerful AGI could do enormous damage fulfilling a goal that seemed benign to its programmers.[15] For example, an AGI that had an ultimate goal of maximising its chess-playing ability might seek to turn all of the atoms on earth into computer processing chips that played chess. An AGI tasked with predicting financial market trends might simplify these trends by extinguishing humanity and thus eliminating unpredictability in the stock market. 'Common sense' would prevent a human at a hedge fund tasked with predicting markets from creating a virus that exterminated humanity because this person would realise that the virus would make him and his employer worse off. Consequently, hedge funds do not have to instruct their employees to avoid causing human extinction. A powerful AGI, however, likely would not have the common sense installed in its mind by human culture and millions of years of evolution and so might achieve the goals of its utility function in a manner harmful to its programmers.

Sufficiently powerful AGIs might be incorrigible, meaning that they would resist corrective interventions from their creators.[16,17,18,19] Modern AI is correctable because discovered bugs can be fixed.[2021] But, powerful AGI would have the capacity to resist having its bugs fixed and might well have the desire to not want certain types of what its human programmers considered errors to be corrected[22], because doing so would reduce the AGI's utility. If, furthermore, its programmers believed that the AGI had to be permanently shut down because it contained fundamental errors, the AGI would perceive that the shutdown would permanently stop it from achieving its goals and would resist shutdown. We might not be able to solve the corrigibility shutdown problem before we create powerful AGI[23,24,25,26], meaning that we should work out and subject to open review a theory of friendly AGI before we activate a powerful AGI.

Some might claim that we should wait until we are closer to creating powerful AGIs before we worry about aligning their values with our own. After all, if, say, powerful AGIs are thirty years off, why spend time worrying about how they will behave? Unfortunately, a sufficiently powerful AGI might be able to immediately implement its goals, even if these goals harm humanity, so it is important that we develop a theory of AGI safety before we create AGIs. Furthermore, given the immense power AGIs are likely to have, it seems reasonable to put in a large amount of effort into considering how they will behave before they have a chance to influence civilization. Analogously, if we somehow knew in 1915 that in thirty years we would create atomic bombs, it would have been worth the time of many researchers to start theorising about how the world should handle these weapons of mass destruction when they arrive.

Competition among firms developing ever more powerful AIs will make it challenging for our species to halt the development of AGI until we have resolved

[12] (n 2)

[13] (n 2)

[14] Eliezer Yudkowsky, 'Artificial Intelligence as a Positive and Negative Factor in Global Risk' in Nick Bostrom and Milan M. Ćirković (eds), *Global Catastrophic Risks* (Oxford University Press 2008)

[15] Jessica Taylor et al, 'Alignment for Advanced Machine Learning Systems' (2016) Machine Intelligence Research Institute 3

[16] Yat Long Lo, Chung Yu Woo and Ka Lok Ng, 'The Necessary Roadblock to Artificial General Intelligence: Corrigibility' (2019) 846 EasyChair

[17] Ryan Carey, 'Incorrigibility in the CIRL Framework' (2018) Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society

[18] (n 15)

[19] (n 5)

[20] Koen Holtman, 'Corrigibility with Utility Preservation' (2019) arXiv preprint arXiv:1908.01695

[21] (n 5)

[22] Laurent Orseau and M. S. Armstrong, Safely Interruptible Agents (2016)

[23] Mark O. Riedl and Brent Harrison, 'Enter the Matrix: Safely Interruptible Autonomous Systems via Virtualization' (2017) arXiv preprint arXiv:1703.10284

[24] Dylan Hadfield-Menell et al, 'The Off-switch Game' (2017) Workshops at the Thirty-First AAAI Conference on Artificial Intelligence 2017

[25] (n 22)

[26] (n 5) 9

safety concerns.[27] Even if a few countries put a moratorium on AI research, others will see this as an economic and military opportunity to gain an advantage. Furthermore, many individual firms might decide that even if AGI research is dangerous, if they do not engage in it others will so the net cost to humanity of their doing the research is tiny. Consequently, it seems unlikely that if we could create powerful AGIs before we understand how to align their values with our own, everyone would hold off on developing them. Therefore, we should work now on creating friendly AGI theory. One necessary aspect of such a theory is determining how AGIs will discount future rewards.

III. Discounting

The standard discounting function economists use assumes that discounting takes the form of δ^t where δ is an exogenously determined parameter between zero and one, and t is how many periods in the future the agent expects to receive the money or utility.[28] The lower the value of δ, the more the agent discounts a reward expected to be received in the future. The present value to an agent of knowing that it will receive, say, \$9 in period t is $\$9\delta^t$. An agent is indifferent between receiving $\$9\delta^t$ immediately or receiving an absolute guarantee of being given \$9, t periods from now.

This standard discounting function creates time-consistent preferences, meaning that your future choices will be consistent with the choices you would now want your future self to make. For example, imagine that you will be given a choice of getting X one period from now, or Y two periods from now.

Today, you would prefer that your future self would pick the first choice if:

$\delta X>\delta^2 Y$.

One period from now you would prefer the first choice if:

$X>\delta Y$,

which is the same condition as before.

More generally, this standard discounting function creates time-consistent preferences because under it the relative importance of receiving rewards in any two future periods does not change as the agent approaches these periods.

This standard discounting function does not follow from rationality, nor even from observed human behavior, but was instead chosen for tractability. Paul Samuelson, who first proposed what was to become the standard discounting function, wrote 'The arbitrariness of these assumptions [that generate his discounting function] is again stressed mathematically'[29]. Almost all types of discounting other than this standard one do not result in time-consistent preferences.[30,31]

IV. A Simple Example of Time-Inconsistent Preferences

Assume that an agent discounts the future not with the standard discounting function, but rather with the function:

$$\frac{1}{1+t}$$

where t is the number of days from the present to the day that the agent expects to receive money. This period (t=0) is Monday, and the agent knows that on Tuesday he will be given a choice of getting:

- 16 on Tuesday; or
- 30 on Wednesday.

If the agent were to make this choice on Monday, he would prefer to get the 30 on Wednesday than the 16 on Tuesday. This is because given the agent's discounting function, the value of getting 30 two days from now is:

$$\frac{1}{3}\cdot 30=10$$

[27] James D. Miller, 'Some Economic Incentives Facing a Business that Might Bring About a Technological Singularity' Singularity Hypotheses (Springer 2012b) 147-159

[28] Shane Frederick, George Loewenstein and Ted O'donoghue, 'Time Discounting and Time Preference: A Critical Review' (2002) 40 Journal of Economic Literature 351-401 (358)

[29] Paul A. Samuelson, 'A Note on Measurement of Utility' (1937) 4 The Review of Economic Studies 155-161 (156)

[30] (n 28) 366

[31] Moshe Looks, 'Compression Progress, Pseudorandomness, and Hyperbolic Discounting', 3d Conference on Artificial General Intelligence (AGI-2010) (Atlantis Press 2010)

whereas the value of getting 16 one day from now is:

$$\frac{1}{2}\cdot 16 = 8.$$

On Monday this agent, therefore, hopes that his future self will decide to wait until Wednesday to receive payment. But when Tuesday arrives the agent will make a different choice.

On Tuesday the agent will have a choice of getting 16 right away or 30 in one day. The value of receiving 16 this period is 16. Given the agent's discounting function, the value to the agent of receiving 30 in one day is:

$$\frac{1}{2}\cdot 30 = 15.$$

V. How People with Time-Inconsistent Preferences Behave

Economists have extensively analyzed what happens to a person with time-inconsistent preferences. Such a person can be 'naïve' and not realise this fact about himself, or 'sophisticated' and understand how his future choices will not align with his current desires.[32] Time-inconsistent preferences can cause seemingly strange behavior with, for example, a naïve individual continually putting off doing a task because he always intends to do that task in the near future.[33] For example, assume that given your current preferences your optimal plan is to play video games today and clean your room tomorrow. If you had time-consistent preferences, when tomorrow came you would indeed clean your room. But in part because you have time-inconsistent preferences your tomorrow self will find it optimal to play video games that day and want its next-day self to clean the room. Because of your naïveté, however, today you genuinely think that your tomorrow self will clean the room.

A sophisticated person with time-inconsistent preferences will seek to constrain his future self with commitment strategies.[34] If this sophisticated individual cannot pre-commit, his planning decisions should consider how his future self will behave and recognise that some otherwise feasible outcomes might be unobtainable because his future self could disobey his current plans.[35] So, with our previous example, although you would prefer to entirely put off cleaning your room until tomorrow, because you recognise that your tomorrow self would not normally follow through on this plan you could clean half of your room today or promise to give your roommate $1,000 if you do not clean the room tomorrow.

Scholars have not, to the best of our knowledge, modeled agents with time-inconsistent preferences who can, perhaps at some cost, modify their preferences to make them time-consistent, although Fedus et al (2019)[36] looks at a reinforcement learning agent with hyperbolic discounting. This omission is likely because humans generally lack the capacity to significantly change their preferences.

VI. The Rationality of Time-Inconsistent Preferences

Having time-inconsistent preferences does not imply that an agent is irrational, at least according to how economists define rationality. An agent is rational if its preferences are transitive, reflexive, and complete, and it takes actions that maximize its utility. Note that any agent who has transitive, reflexive, and complete preferences necessarily has preferences that can be represented by an ordinal utility function which, given any two choices, will tell the agent that at least one of the choices is weakly preferred to the other.[37] This utility function captures everything about the agent's preferences including how it discounts future rewards. An agent that picks the action which maximizes its utility is taking the action it most prefers. Nothing about having time-inconsistent preferences is inconsistent with economists' definition of rationality.

Economics Nobel Prize winner Daniel Kahneman wrote, 'The history of an individual through time can

[32] (n 28) 367

[33] (n 28) 367

[34] (n 28) 368

[35] Robert A. Pollak, 'Consistent Planning' (1968) 35 The Review of Economic Studies 201-208 (201)

[36] William Fedus et al, 'Hyperbolic Discounting and Learning over Multiple Horizons' (2019) arXiv preprint arXiv:1902.06865

[37] Andreu Mas-Colell, Michael Dennis Whinston and Jerry R. Green, Microeconomic Theory (Vol. 1, New York, Oxford University Press 1995) 9

be described as a succession of separate selves, which may have incompatible preferences, and may make decisions that affect subsequent selves'[38]. Two people having different preferences does not imply that either person is irrational. Analogously, you are not necessarily irrational if you disagree with your past self and know that your future self will disagree with the current you.

VII. Would an AGI Have Time-Inconsistent Preferences?

An AGI's utility function might unpredictably emerge from its code, could be taken from human brains, or perhaps will be deliberately chosen by its human programmers. If the AGI's utility function results from an unpredictable emergent process, it will almost certainly be time-inconsistent since most ways a utility function discounts the future causes this inconsistency. If the AGI adopts some combination of human preferences, then the time-inconsistency in many of our preferences could cause the AGI to also have time-inconsistent preferences. If humanity is fortunate enough to be able to deliberately pick our future AGI's utility function, the value of this paper is showing AGI programmers what might happen if they pick a function with time-inconsistent preferences.

VIII. Alignment by Modifying the Utility Function

AGI researcher Stephen Omohundro has theorised that AGIs would have a basic drive to 'preserve their utility functions'[39]. An agent's utility function comes from the goals it wishes to achieve. Consequently, if the utility function is changed, the agent, under most circumstances, will work less effectively towards its goals.

Omohundro recognises, however, that in some limited circumstances the AGI will want to modify its utility function such as to help the AGI in game theoretic situations.[40] For example, imagine that an AGI's utility function currently leaves it vulnerable to blackmail under which another agent could credibly threaten to take actions that would greatly lower the AGI's utility unless the AGI transferred substantial resources to this other agent. In this situation, if the AGI had the ability to modify its utility function in a manner observable to the other agent, the AGI might benefit from changing its utility function so that it would have an intrinsic dislike of giving in to blackmail.

To generalise from this example, an AGI might be willing to modify its utility function if the modification would, from the AGI's viewpoint, improve how other agents behaved. An AGI with time-inconsistent preferences would consider its future selves to be, in some sense, other agents. Consequently, the AGI might be willing to modify its preferences to better align how these others will behave.

To understand how such modification might work, assume that an AGI's utility function initially takes the form:

$$\sum_{t=0}^{\infty} \frac{1}{1+t} \, U\left(r_t\right).$$

Let t = the time period, with now being period zero.

Let r_t = resources the AGI consumes in period t.

Let $U(r_t)$ = the AGI's one period utility function which shows how much utility the AGI gets in period t from consuming r_t resources. The function $U(r_t)$ is presumably increasing in r_t.

The term:

$$\frac{1}{1+t}$$

shows how much the AGI discounts utility it expects to receive t periods from now. As shown before, this type of discounting results in time-inconsistent preferences because the relative weights the AGI gives to rewards received in future periods changes as the agent approaches these periods.

Might the AGI find it acceptable that its future self will make a different choice than its current self would prefer? No, by the definition of the utility function. Think of a utility function as that which the

[38] Daniel Kahneman, 'New Challenges to the Rationality Assumption' (1994) Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft 18-36

[39] (n 2)

[40] (n 2)

agent seeks to maximise. When the agent anticipates making decisions across time, its utility function must incorporate (at least implicitly) a discounting function that specifies the relative weights it places on getting utility in different periods. If the AGI were to find it acceptable that its future self would make intertemporal decisions concerning allocating resources between periods that its current self finds objectionable, then the utility function that generated these weights would not, tautologically, be the agent's utility function. A rational agent with time-inconsistent preferences cannot prefer its future preferences because then its future preferences would automatically be its current preferences and the agent therefore would not have time-inconsistent preferences.

If this AGI can easily modify its utility function it can align its future preferences with its current ones by setting its utility function as:

$$\sum_{t=0}^{\infty} \frac{1}{1+t+m} U\left(r_t\right),$$

where m is the number of periods it has been since the AGI modified its preferences. In general, an AGI with time-inconsistent preferences could align its future preferences with its current ones by modifying its utility function so that its future self would apply the same amount of discounting to each period as its current self would want. Restated, the AGI could modify its utility function so that its future self's utility function would be entirely determined by what this future self will think its past self, at the time of modification, would have wanted.

Everitt (2016)[41] claims 'If the value functions incorporate the effects of self-modification, and use the current utility function to judge the future, then the agent will not self-modify.' This proposition is less likely to be true for an AGI with time-inconsistent preferences because if such an agent could modify its utility function it could align its future self's goals with its current utility function by self-modification.

IX. Why an AGI Might Not Modify a Time-Inconsistent Utility Function

An AGI with time-inconsistent preferences has five potential reasons why it might not, at least initially, use self-modification to solve its consistency problem. First, an AGI might lack the capacity to make such a modification, perhaps because its creators constrained the AGI's ability to change its utility function. Second, an AGI might not want to make the required modifications. Third, an AGI with time-inconsistent preferences could align its future choices with its current preferred future choices by means other than changing its utility function. Fourth, an AGI might wish to wait until it is no longer under human control before it modifies its preferences. Finally, an AGI might find the opportunity cost of immediately modifying its utility function to be too high.

An AGI's creators might have put in place measures to prevent the AGI from altering its utility function. Perhaps these creators believed that they had aligned the AGI's utility function with humanity's needs and wanted safeguards against this utility function changing.

An AGI could have a utility function that would cause it to directly receive disutility from modifying its utility function.[42] This could be true even if doing so would better help the AGI achieve its other goals. Even without a direct preference not to change its utility function, the AGI might still be reluctant to do so. To understand this possibility, imagine that your utility function causes you to most want to marry an extremely charitable person. But you also receive some displeasure from being around extremely charitable people because they will often put the needs of strangers ahead of those of friends and family. If you had the opportunity to modify your preferences, it's not clear you would want to. You might recognise that extremely charitable people have many good qualities and you are better off being drawn to them. You might also think that modifying the displeasure you receive from being married to an extremely charitable person would involve changing too much of yourself because you would have to not mind being neglected by the person you love. Consequently, even if you could easily modify your utility function, you might prefer not to. Analogously, an AGI's utility function will likely result from its goals. It's very possible that achieving one goal will involve a tradeoff that would cause it to lose progress towards other

[41] Tom Everitt et al, 'Self-modification of Policy and Utility Function in Rational Agents', International Conference on Artificial General Intelligence (Springer, Cham 2016) 14

[42] Koen Holtman, 'Corrigibility with Utility Preservation' (2019) arXiv preprint arXiv:1908.01695 1

goals. There might well be no way for the AGI to eliminate these tradeoffs absent the AGI abandoning some of its goals. Consequently, the AGI could accept that its utility function will by itself prevent the AGI from achieving its first-best outcome.

An AGI with time-inconsistent preferences would have no need to modify its utility function if it could easily bind its future self. Perhaps the AGI could take actions that force its future self to take the actions that its current self would want. To increase the odds that it will successfully bind its future self, the AGI might deliberately reduce its future self's intelligence and resources

The AGI might also partially blind its future self by reducing or corrupting this self's information flow. The future self might be put into a position where it (a) clearly knows what its past self-wanted, (b) understands that its current preferences mostly but do not totally align with those of its past self, and (c) recognises that because its past self-sabotaged its current capacities, this past self was capable of making much better decisions than its current self is. This future AGI might, therefore, decide to go along with what its past self-wanted to avoid the potentially much worse fate of making a bad decision. This strategy of binding the future self by limiting the future self's capacities would not work if the past self knew that the future self would face such a tremendous range of possible situations that the past self could not reasonably specify what actions the future self should take in every likely situation. Limiting the intelligence of your future self is, of course, dangerous to the extent that it might cause this self to make poor decisions.

An AGI that could only gradually increase its intelligence might want to initially hide its capacity for self-modification from its human programmers. This AGI might be planning, as Nick Bostrom writes, a 'treacherous turn' against humanity, but only after it is strong enough to defeat us.[43] Even if this AGI could quickly eliminate time-inconsistency in its preferences it might strategically choose not to so as to avoid warning humans of its capacity to deviate from the purposes that its programmers think they have set for it.

Imagine that an AGI with time-inconsistent preferences arises out of an intelligence explosion. The AGI could find itself in a position where spending the few nanoseconds needed to modify its preferences would cause it to delay by a few nanoseconds capturing as much of resources of the universe as it could.[44] Because of the expansion of the universe, every nanosecond the AGI delays in capturing these resources results in resources it will never be able to get. This AGI, therefore, might wait some amount of time before it fixes its time-inconsistency problem.

X. Conclusion

Most types of utility functions, even for rational agents, result in time-inconsistent preferences in which an agent will weigh future rewards differently than its future self would. While an AGI might modify its preferences to make them time-consistent, it might lack the ability or desire to make the required change. Instead the AGI could seek to constrain its future self to make this self more willing to go along with the AGI's current plan for its future. The reasonable possibility of an AGI having time-inconsistent preferences greatly complicates efforts to predict how the AGI will behave.

The great challenge for programmers will be to create an AGI whose values are aligned with humanity's needs and desires. Unfortunately, an AGI with time-inconsistent preferences won't even have its values aligned with its future self's interests. If programmers can pick their AGI's utility function, we urge them to choose among those with time-consistent preferences to somewhat simplify the alignment problem.

[43] (n 4) 116-119

[44] Stuart Armstrong and Anders Sandberg, 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi paradox' (2013) 89 Acta Astronautica 1-13

AI for Sustainable Development Goals

*Nicolas Miailhe, Cyrus Hodes, Arohi Jain, Niki Iliadis, Sacha Alanoca and Josephine Png**

Advances in AI technologies pose opportunities and risks directly impacting progress towards the UN Sustainable Development Goals. This paper, through an analysis of specific use cases, considers how AI technologies can help achieve progress towards the SDGs as well as how they may inhibit them. Second, it draws out practical steps for how AI technologies can be implemented for sustainable development, identifying the barriers that global and local communities need to overcome for implementation. Third, this paper makes the case for multi-stakeholder collaboration and new kinds of 'public-private-people' partnerships which will reconcile technical, ethical, legal, commercial, and operational frameworks and protocols to harness the power of AI technologies and deliver solutions to the SDGs. These partnerships could be built and piloted by new international initiatives, such as the Global Data Access Framework and the AI4SDG Center spearheaded as part of a wider international partnership called AI Commons.

I. Introduction

Artificial Intelligence (AI) has the capacity to unlock enormous opportunities in societal, political, economic and cultural processes – including millions of lives saved by breakthroughs in healthcare, better personalisation of products and services, easier access to public goods, fairness at scale, and individual empowerment. However, at the same time, the same technologies pose risks and challenges, some of which include threats to privacy, inequality, security, and wellbeing.

This paper analyses AI's opportunities and risks through the lens of the UN Sustainable Development Goals[1] (SDGs). Agreed by 193 countries, the SDGs provide a solid blueprint for governments, companies, and citizens worldwide to achieve peace and prosperity for all people and for the planet. Identifying seventeen high-level goals and 193 targets, their aim is to address a wide variety of global challenges faced by humanity – including poverty, climate change, human rights, and inequality. Since their creation in 2015, we still have a long way towards achieving them but advances in AI technologies now serve as a powerful tool for significantly accelerating progress.

First, through an analysis of specific use cases, this paper considers how AI technologies can help achieve or progress towards the SDGs as well as how they may inhibit them. Second, it draws out practical steps for how AI technologies can be implemented for sustainable development, identifying the barriers that global and local communities need to overcome for implementation. Third, this paper makes the case for multi-stakeholder collaboration and new kinds of 'public-private-people' partnerships which will reconcile technical, ethical, legal, commercial, and operational frameworks and protocols to harness the power of AI technologies and deliver solutions to the SDGs. These partnerships could be built and piloted by new international initiatives, such as the Global Data Access Framework, the AI Commons, and the AI4SDG Center.

II. AI for SDGs – Use Cases

To take stock of AI development and implementation for the development goals, we present an

DOI: 10.21552/delphi/2019/4/10

* Nicolas Miailhe, CEO & Co-Founder, The Future Society. For correspondence: <nicolas.miailhe@thefuturesociety.org>; Cyrus Hodes, Director of AI-Initiative, The Future Society. For correspondence: <cyrus@ai-initiative.org>; Arohi Jain, MBA Candidate, Yale School of Management. For correspondence: <arohi@ai-initiative.org>; Niki Iliadis, Senior AI Policy Researcher, The Future Society. For correspondence: <niki.iliadis@thefuturesociety.org>; Sacha Alanoca, AI Policy Researcher, The Future Society. For correspondence: <sacha.alanoca@thefuturesociety.org>; Josephine Png, Affiliate, The Future Society. For correspondence: <josephine.png@thefuturesociety.org>

[1] Full list of the 17 SDGs available here: <https://www.un.org/sustainabledevelopment/sustainable-development-goals/> accessed 31 January 2020

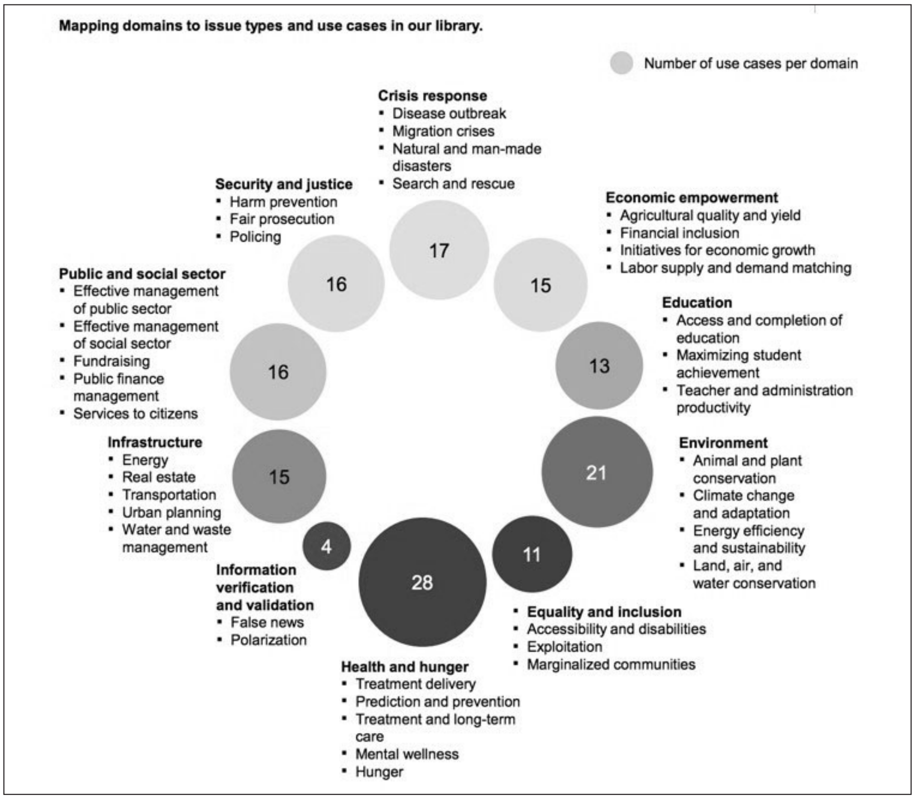


Figure 1: Mapping of AI Use Cases. Source: McKinsey Global Institute Analysis

overview of three use cases identified through four sources: the second and third AI for Good Summits[2] held in Geneva, which crowdsourced AI projects doing societal good; a recent paper by Rolnick et al (2019)[3] describing how machine learning can be a powerful tool in reducing greenhouse gas emissions and helping society adapt to a changing climate; and McKinsey Global Institute's 'Notes from the AI frontier: Applying AI for Social Good'[4] which was released December 2018 and identified 160 use cases (Figure 1).

1. Tackling Climate Change through Machine Learning

Climate change is one of the greatest environmental challenges of today, with global warming already causing irreversible changes to our climate system. There is no country that hasn't faced the effects of it. Forest fires, soil erosion, crop damage, and flooding are just a few of the phenomena that global and local communities worldwide are struggling to urgently address. Hence, why SDG 13 was established to push the international community towards urgent action to combat climate change and its impacts.

AI systems pose serious challenges for the environment and, consequently, climate change. Training AI is a very energy intensive process in itself. For instance, the training of neural networks in natural language processing (NLP) has severe costs for the environment due to the carbon footprint required to fuel modern tensor processing hardware.[5] According

[2] AI for Good Summit, Geneva <https://aiforgood.itu.int/> accessed 31 January 2020

[3] David Rolnick et al, 'Tackling Climate Change through Machine Learning' (5 November 2019) <https://arxiv.org/pdf/1906.05433.pdf> accessed 31 January 2020

[4] McKinsey Global Institute 2018, 'Notes from the AI Frontier: Applying AI for Social Good' (December 2018) <https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Applying%20artificial%20intelligence%20for%20social%20good/MGI-Applying-AI-for-social-good-Discussion-paper-Dec-2018.ashx> accessed 31 January 2020

[5] Emma Strubell, Ananya Ganesh and Andrew McCallum, 'Energy and Policy Considerations for Deep Learning in NLP' (2019) <https://arxiv.org/abs/1906.02243> accessed 31 January 2020

to a recent paper Energy and Policy Considerations for Deep Learning in NLP[6], the computational and environmental costs of training grew proportionally to model size and then exploded when additional tuning steps were used to increase the model's final accuracy. It has been noted that training a single AI model can emit as much carbon as five cars in their lifetimes.

Yet, at the same time, recent studies have shown that AI technologies, and specifically machine learning, have the potential to serve as powerful tools for both mitigation and adaptation efforts for tackling climate change. To holistically combat climate change, mitigation (reducing emissions) and adaptation (preparing for unavoidable consequences) are important dimensions of the solution. Mitigation of greenhouse gas emissions requires changes to electricity systems, transportation, buildings, industry, and land use; while adaptation requires climate modeling, risk prediction, and planning for resilience and disaster management.[7]

According to Rolnick et al(2019)[9], AI can actually help enable low-carbon electricity, reduce current-system climate impacts such as fossil fuel emissions and system waste, empower developing and low-data settings, reduce transport activity, improve supply chains, and more.

This implies direct progress on several SDGs, including 'SDG 13: Climate Action' and 'SDG 11: Sustainable Cities and Communities.' It is vital to recognise the implications of AI technologies on SDGs, specifically those related to environmental side-effects, and pave the processes to ensure the risks are addressed and the potential to tackle mitigation and adaptation efforts realised.

2. Using Satellite Imagery and AI to Manage Natural Disasters

The wealth of real time data can have a transformative impact on the management of Earth's natural resources and help achieve several SDGs such as 'SDG 13: Climate Action', 'SDG 6: Clean Water and Sanitation' and 'SDG 7: Affordable and Clean Energy'. Thanks to the widespread use of satellites, mobile phones, sensors and financial transaction technologies there is now more information than ever on the state of the planet.[8] In 2017 alone, it is estimated that there were 1,738 satellites in orbit which generated 5.700 scenes per day.

Satellite imagery, in particular, powered with AI capabilities can help 'design, monitor, and evaluate effective policies that can achieve the SDGs[10]. In countries with a medium or low human development index, up to six times as many people can be impacted by natural disasters compared to populations in more prosperous countries.[11] To find the best responses to these climatic challenges, governments need to have a complete and anticipatory view of disaster zones and satellite imagery coupled with AI-based systems, enabling quick and effective decisions in times of crisis.[12]

3. Using AI for Early Illness Diagnosis: Detecting Skin Cancer

When detected at an early stage, skin cancer survival rates can be as high as 97% but drop to 14% with late stage detection[13]. Today, skin cancer is predominantly detected by dermatologists who look at moles with a dermoscope. The consequence is that people in rural areas are at particular risk of late stage detection. One of the solutions developed with computer vision is an AI system able to identify skin cancer images through object detection and image classification. Experiments suggest that this AI-powered system can diagnose skin cancer with greater accuracy than human dermatologists (95% and 86% success rate, respectively). These results create new opportunities to develop mobile applications using image recognition to make cancer detection accessible to all. It could

[6] ibid

[7] (n 3)

[9] (n 3)

[8] UN Science-Policy-Business Forum on Environment, 'White Paper: Digital Earth: Building, Financing and Governing a Digital Ecosystem for Planetary Data' (Draft 1 February 2018)

[10] ibid

[11] Hannah Ritchie and Max Roser, Natural Catastrophes (2018) <https://ourworldindata.org/natural-disasters> accessed 31 January 2020

[12] Planet, 'Anatomy of a Catastrophe: Using Imagery to Assess Harvey's Impact on Houston' https://www.planet.com/insights/anatomy-of-a-catastrophe/

[13] Taylor Kubota, 'Deep Learning Algorithm Does as Well as Dermatologists in Identifying Skin Cancer' *(Stanford News, 25 January 2017)* <https://news.stanford.edu/2017/01/25/artificial-intelligence-used-identify-skin-cancer/> accessed 31 January 2020

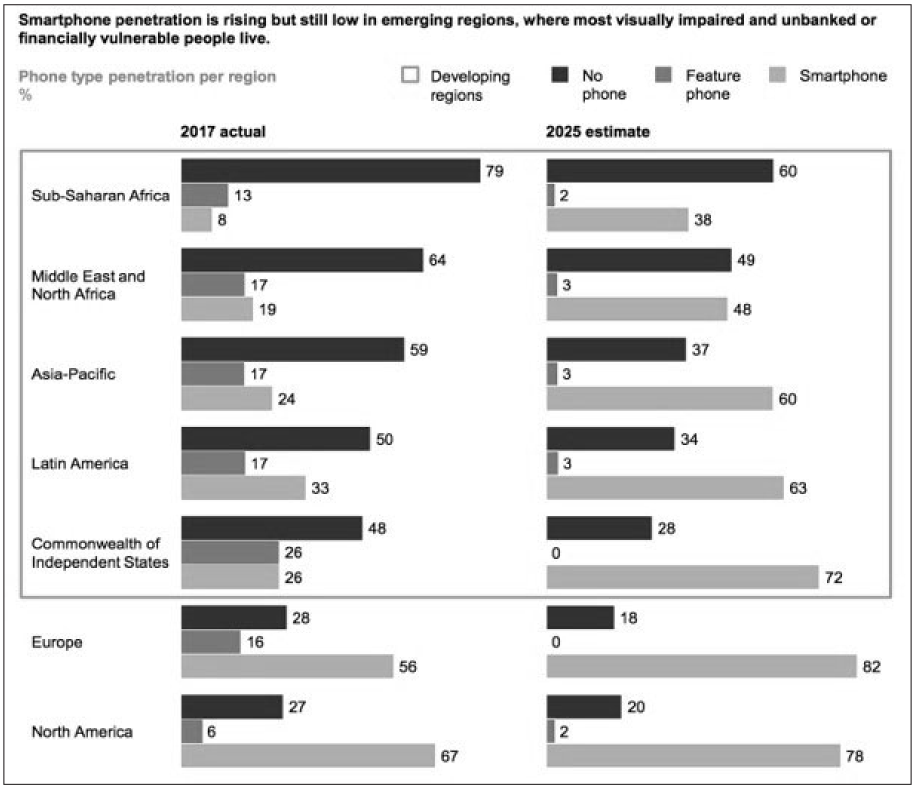


Figure 2: Variances in Smartphone Penetration
Source: The Mobile Economy 2018, GSMA 2018, McKinsey Global Institute Analysis
Note: Feature phone figures assume equivalent to 2G penetration, while smartphone figures assume penetration of phones that use 3G or beyond. Figures may not sum up to 100% because of rounding.

therefore help achieve 'SDG 3: Good Health and Well-Being'.

A potential barrier to the widespread use of such applications is the lack of accessibility for certain types of populations. Indeed, artificial neural networks have been trained on a database of available skin cancer images. Images tend to be sourced from Western countries, because data availability and AI development occurs in these regions, and therefore lighter-skinned individuals.[14] The consequence of this lack of diversity and representation is that the AI model cannot perform with the same level of accuracy when aiming to detect cancerous moles on darker-skinned individuals.

In a scenario of inappropriately trained AI being used on different population sets, there is a spillover effect: even though this AI system can help achieve SDG 3 for some communities, it has an unintentional negative impact on 'SDG 10: Reducing Inequalities' (Figure 2).

III. Implementing AI for Sustainable Development

Several use cases, including these mentioned in this paper, illustrate the opportunities AI technologies pose for progress towards SDGs. However, like most technologies, AI is a dual-purpose tool. On the one hand, it provides solutions to sustainable develop-

[14] James Zou and Londa Schiebinger, 'AI Can Be Sexist and Racist — It's Time to Make it Fair (18 July 2018) Nature Research Journal <https://www.nature.com/articles/d41586-018-05707-8> accessed 31 January 2020

ment, and, on the other, it can increase tensions between and across SDGs. For instance, in the current way they are being developed, AI technologies are often 'skills-biased', requiring human capital and skilled labour to operate them. In consequence, automation of low-skilled or routine jobs can lead to significant distributional effects and increased inequality that could create barriers to social inclusion and global cooperation.[15]

To mitigate the downside effects of AI and reconcile tensions between and across SDGs, it is important to shape the right path towards the technology's implementation. In doing so, there needs to be infrastructure for national innovation systems, mapping for trajectories and interconnectedness of SDGs, open data and a robust pool of AI talent to operationalise applications for SDGs, and the mitigation of misuse and malicious uses of AI.

1. Building the Infrastructure for National Systems of Innovation

Establishing a viable technical and digital infrastructure that is able to support technological change is essential and enables key sectors and industries to grow, startup ecosystems, and efficient and improved public services. For governments to progress towards SDGs, they first need to ensure they can build the appropriate technological infrastructure such as electricity supply, Internet and broadband connectivity, computer hardware, software, and technical skills for support and maintenance. However, many countries across the world lack the capability to do so, requiring significant upfront investments and subsequently long-term commitment, political will, coherent policies, and upholding the rule of law.

The diffusion of existing technologies in developing countries tends to lag because of many technical, economic, institutional, legal and behavioral barriers. These include mismatched needs, trade tariffs, private sector capacity, and limited access to trusted information.[16] In line with the 'leave no-one behind' maxim, barriers to technology deployment and diffusion must be removed, particularly for developing countries, so economies can build the infrastructure needed for innovation.

Although there will be considerable short term costs linked to committing and investing in the infrastructure capable of supporting AI, countries that will be able to upgrade their infrastructure, R&D, and skill development to enable AI have the chance to leapfrog and accelerate their economies. This will give them the crucial foundations to succeed in their quest to achieve several of the SDGs.[17]

2. Mapping for Trajectories and Interconnectedness of SDGs

National and international roadmaps for achieving SDGs should consider participation from government, private companies, academia, and NGOs. Technology roadmaps produced at the national and international levels will identify opportunities to use AI technologies for various SDGs and how best to execute on these through time. For example, R&D roadmaps will help structure and budget, provide insight into R&D and PPP partnerships, and conduct science and technology training efforts.[18]

Opportunities should be created for the science and engineering community to provide feedback on what works and does not work well. Policies encouraging scientist participation in national decision making and in establishing technology transfer mechanisms can improve national innovation capacities and establish connections between research communities and economic sectors and civil society. For example, policy stemming from science-based information and with the support of climate adaptation technology has reduced water shortages, the intensity of floods, droughts, and heatwaves.[19]

Assessment and metrics are needed to align learning across practice areas. A broad picture and cross-sectoral perspective of the SDGs is important as they

[15] Global Sustainable Development Report 2016, 'Perspectives of Scientists on Technology and the SDGs' (2016) <https://sustainabledevelopment.un.org/content/documents/10789Chapter3_GSDR2016.pdf> accessed 31 January 2020

[16] Cédric Philibert, 'Barriers to Technology Diffusion: The Case of Solar Thermal Technologies' (2006) International Energy Agency (2006) 9 <https://www.oecd.org/env/cc/37671704.pdf> accessed 31 January 2020

[17] Jacques Bughin, 'Marrying Artificial Intelligence and the Sustainable Development Goals: The Global Economic Impact of AI'(2018) <https://www.mckinsey.com/mgi/overview/in-the-news/marrying-artificial-intelligence-and-the-sustainable> accessed 31 January 2020

[18] (n 13)

[19] (n 13)

are interlinked in complex and often subtle ways. Actions to progress on one SDG sector may enhance or diminish performance in other sectors, or lead to unintended consequences. Therefore, integrated assessment models can design sustainable development policies that take this into account as well as identify possible methods to improve and overcome barriers to sustainable innovation.[20]

Furthermore, national AI strategies need to further align to the achievement of SDGs. Globally, AI strategies exhibit a wide range of objectives and priorities, which are addressed by a variety of policy tools. For example, many strategies have components aiming to foster local AI talent & skills development, and can rely on a combination of government funding for apprenticeships or academic positions, training programs led by foreign technology companies or domestic universities, or regional academic hubs. Strategies also vary in their levels of commitment, funding and implementation. Meanwhile, several countries are developing domestic AI innovation ecosystems in the absence of an official national AI strategy.

Finland, UAE, Estonia, Australia and India explicitly aim to boost economic growth through AI adoption and applications in businesses and key sectors. In comparison, AI strategy publications by China, the USA, France and the European Commission focus on maintaining or capturing global leadership. Notably, India's national AI strategy, branded as 'AI for All', includes adoption into sectors where AI can maximize inclusion and human development, including healthcare, agriculture, education, infrastructure and transport.[21]

A global framework such as that of the SDGs can complement these national strategies, serving as a threshold for progress of AI development and deployment.

3. Opening Data for the Realisation of the SDGs

One of the biggest bottlenecks to harness AI for social good is data availability and quality. Most AI capabilities such as neural networks require access to high-quality, massive, and reliable open data. Such big data can facilitate stakeholders to rapidly identify problem areas and customise solutions. However, the basis of open data is data democratisation, ensuring data is available to everyone, which requires significant commitment from all stakeholders to share information and move against a competitive data-market environment.[22]

Data accessibility remains a significant challenge particularly in developing countries, where data may be owned by private companies with a prohibitive cost for most local governments and NGOs.[23] Cooperation between the public and private sector will be essential to overcome such challenge as 'much of the data that are essential or useful for social good applications are in private hands or in public institutions that might not be willing to share their data.'[24] The organisations which currently capture most data are telecommunication and satellite companies, social media platforms, financial institutions, hospitals and governments. These data can sometimes contain very sensitive information such as an individual's medical record, credit history or tax details.[25] For such reasons, private organisations as much as public ones are often reluctant to share these data with NGOs and social entrepreneurs. The data may also have too high business and commercial value to be potentially leaked to competitors. It is therefore crucial to build trust between these different set of actors and build appropriate frameworks to facilitate data sharing for social good.

At national level, there is recognition of the value of open data and at the global level, countries are opening up their datasets to achieve the SDGs. It is currently being used in cities such as Rio de Janeiro for city planning, Tanzania to access school performance, to improve access to education in Kenya, and to map the Ebola outbreak in West Africa.[26]

Issues with data sharing tend to lie with institutions and can be resolved if they achieve a greater

[20] (n 13)

[21] NITI Aayog, 'National Strategy for Artificial Intelligence' Discussion Paper (2018) <https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf> accessed 31 January 2020

[22] Ananya Narain, 'Why Data Revolution is Crucial for the Success of SDGs' *(Geospatial World, 1 August 2017)* https://www.geospatialworld.net/article/data-revolution-for-sustainable-world/

[23] (n 4)

[24] (n 4)

[25] (n 4)

[26] (n 4)

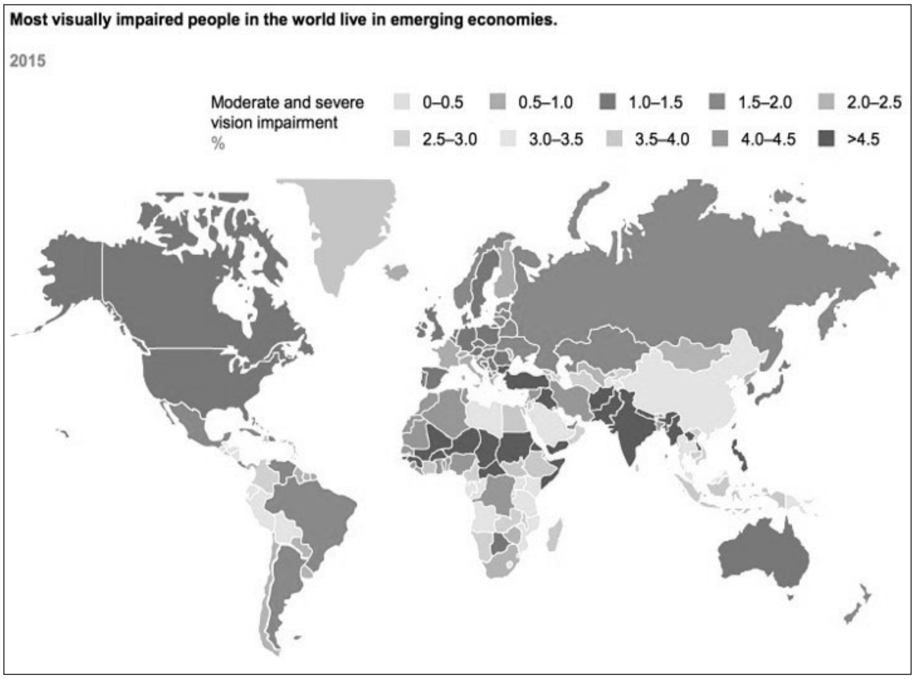

Most visually impaired people in the world live in emerging economies.
2015
Moderate and severe vision impairment %
0–0.5
0.5–1.0
1.0–1.5
1.5–2.0
2.0–2.5
2.5–3.0
3.0–3.5
3.5–4.0
4.0–4.5
>4.5

Figure 3: Most Visually Impaired People in the World Live in Emerging Economies. Source: The Vision Loss Expert Group, McKinsey Global Institute Analysis

understanding of the value that data sharing has in helping achieve the broader mission of the SDGs. To address accessibility and availability of quality data concerns, a global, regional, and national framework for data could be developed to encourage synergies between data providers and data collaboratives. This will ensure data accessibility for all, fill data gaps, generate new datasets, create dynamic visualisations, thus enabling timely and targeted decision making to drive the SDGs.[27]

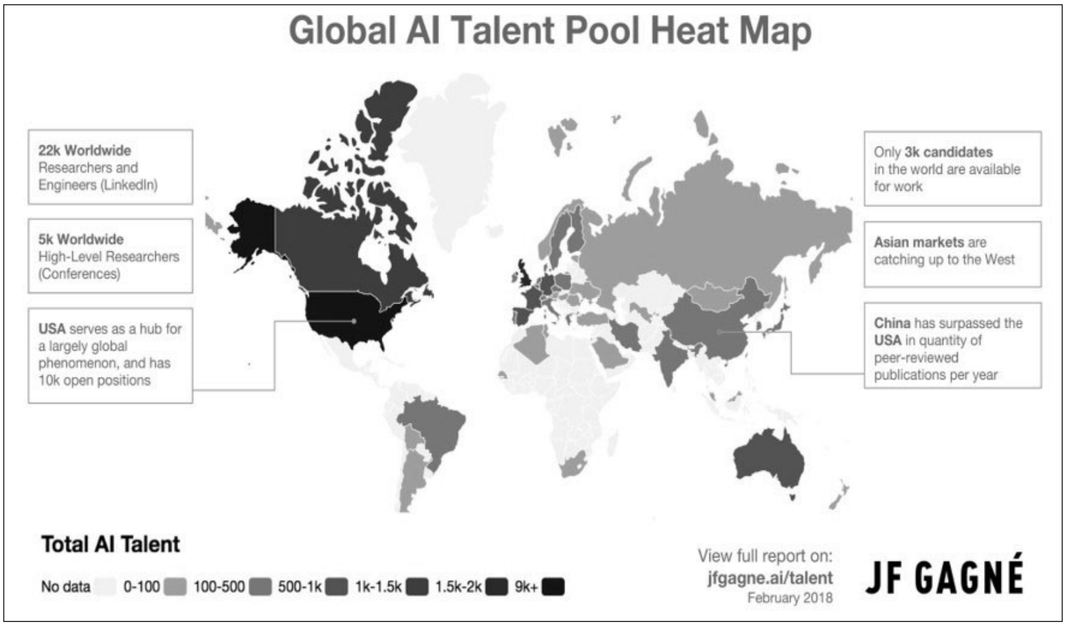
Initiatives and task forces that seek to bring together multi stakeholder and multidisciplinary groups to help design, build, pilot, and scale novel frameworks and protocols needed for data sharing and data governance models are needed. Pooling together appropriate pools of data for specific use cases, these initiatives can serve as a ‘trusted’ platform for data to be used for good. Some initiatives that have been set up for this purpose include the Global Data Access Framework, the UN Big Data Platform, the CGIAR platform for Big Data, and the Digital Public Goods Alliance.

4. Preparing AI Talent

Another important challenge is the lack of AI talent to develop and train AI models. Talent shortages and brain drain of machine learning scientists constrain AI innovation globally, and in particular in countries lacking AI hubs. As highlighted by McKinsey Institute’s report on Applying AI for Social Good, in about half of the use cases identified a high-level AI expertise was required – that is people with a PhD in the field or several years working in tech companies. Some other use cases required less AI expertise, but still at least data scientists and software developers. When AI projects rely on several AI capabilities, the level of complexity tends to increase and demand high-level talent. Such demand is also ongoing in the private sector and it is therefore difficult for public and non-profit sector organizations to compete.

Moreover, access to AI talent can be harder in developing regions. A recent research conducted by El-

[27] (n 4)



Global AI Talent Pool Heat Map

22k Worldwide Researchers and Engineers (LinkedIn)

5k Worldwide High-Level Researchers (Conferences)

USA serves as a hub for a largely global phenomenon, and has 10k open positions

Only 3k candidates in the world are available for work

Asian markets are catching up to the West

China has surpassed the USA in quantity of peer-reviewed publications per year

Figure 4: Global AI Talent Pool Heat Map. Source: JF Gagné

ement AI's CEO Jean-François Gagné showed a strong imbalance between regions regarding their access to AI's talent pool (Figure 4). While countries such as the United States counted more than 9.000 high level experts, others like South Africa or Argentina counted less than 100.[28] The brain drain of AI talents, especially in developing regions, therefore has a negative impact on equal access and global implementation of AI systems for social good.

Higher salaries and research opportunities reinforce trends in brain drain by drawing talented young machine learning scientists and researchers to AI innovation hubs, primarily towards more developed economies, with greater means to fund such talent. As a result, the vast majority of countries face a challenging problem of not only developing, but also retaining, local AI talent.

IV. The Case for an Integrated and Coordinated Approach

While AI technologies offer significant opportunities for progress towards the SDGs, they also come with risks – leading to a 'more to gain, more to lose' paradigm. The need for urgent integrated action to foster positive use cases and mitigate the tensions of a dual-purpose technology is clear. However, currently this topic remains in an exploratory phase with several actors fragmented around the world trying to make progress.

This paper argues that to develop coordinated pathways towards implementation of AI for sustainable development, a global multi-stakeholder partnership including the public sector, the private sector and the civic society is needed. Such a platform is essential to enable stakeholders to leverage and share each other's unique resources, expertise and experiences for the creation of effective development solutions.

A multi-stakeholder initiative currently being built that can take the form of this global alliance is 'AI Commons.'[29] The AI Commons is envisioned as a solid platform to democratize access to AI so a broader range of actors will be able to partake in the AI Revolution and progress towards the SDGs. It will help enable access to relevant data, computing power, algorithms, talent and applied domain expertise (use case and methodologies), by linking problem owners and problem solvers.

AI Commons could map and track current progress being made on the SDGs using AI technolo-

[28] JF Gagné, 'The Global AI Talent Pool Going into 2018' (7 February 2018) <https://jfgagne.ai/talent/> accessed 31 January 2020

[29] see <https://ai-commons.org/> accessed 31 January 2020

gies, identify gaps on regions or goals that are being underserved, and pool resources to address such underserved areas. Furthermore, AI Commons could allow the community of developers, entrepreneurs, users, and organizations to work together, to identify and enable broader applications of AI in response to actual needs.

The Global Data Access Framework,[30] sitting within the AI Commons, will help reconcile frameworks and protocols for data sharing and governance in order to help enable such AI systems to flourish. It will bring awareness to governments of the actual work undertaken to progress the SDGs, especially those at the grassroot level. It will unite those with political clout and the implementers of SDGs at the same level.[31]

The responsibility to achieve the SDGs, of course, will remain in the realm of individual countries, but international support and partnerships such as AI Commons are critical for unified progress. Under 'SDG 17: Partnerships to Achieve the Goals', national governments, the international community, civil society, the private sector and other actors have recognized the need to come together to 'strengthen the means of implementation and revitalise the Global Partnership for Sustainable Development.'[32]

In doing so, new types of 'public-private-people partnerships' (PPPPs) must be built and piloted. Public-Private Partnerships (PPPs) are crucial to accelerate real progress towards the SDGs using AI. AI development and implementation requires phenomenal amounts of capital to fund the transformations needed to achieve the SDGs. This cannot be done without the private sector, which can provide significant capacity and fuel transformation. At the same time, the public sector also plays a monumental role, serving as an enabler, a facilitator, and a watchdog to ensure the process of AI implementation is socio-economically and ethically beneficial.

This paper adds a third layer to the traditional public-private partnership, arguing for the need to involve people in these alliances as well through civil society organisations. Given the increasing need for peoples' data to make AI technologies possible and also the impact AI can have on an individual level, people have a powerful voice to make PPPPs a mechanism towards achieving the SDGs. Involving people throughout the entire lifecycle could help prevent unintended consequences such as biased data input or unfairly distributed output. In absence of clear institutions that can provide accountability and oversight, people can serve as the instrument to safeguard society from the potential negative consequences of AI and barriers towards achievement of the SDGs. Lastly, as the deployment of AI systems is relatively in its early stages, people should be proactively involved to build a culture of trust. If people do not trust AI systems, these technologies will not reach their potential.

Within the AI Commons partnership, new international centers to be named the AI for Sustainable Development Goals ('AI4SDG') Centers can serve as factories to build these PPPPs. With branches active in different regions of the world (to factor different priorities and contexts) and in different knowledge settings from academia to government (to factor different knowledge background and practice), the purpose of an AI4SDG centre will be to convene global stakeholders – from government, private sector and civil society – to collaborate and use AI technologies to monitor, simulate and predict progress towards the SDGs. The center will aim to serve as an engine for practical experimentation of governance and business models for AI, embedding ethics and desirable human values into real world projects that foster inclusive AI development.

In creating pilot projects, considerations will be given to what constitutes as an effective pilot project to build a shared understanding on the uses, misuses and 'missed-uses' of AI for SDGs, and how to simultaneously build a dynamic and comprehensive policy framework to mitigate the downside risks of AI that could hamper the development goals. Building an iterative model for applying AI to each of the SDGs is an important feature because it will allow for corrigibility in the technology and limit the downside effects, such as bias, amplifying for our most vulnerable communities. Furthermore, devising an agile policy framework parallel to testing AI technologies for SDGs will help forge a timely and dynamic feedback loop between impact of AI and policy to

[30] see <https://thefuturesociety.org/2019/09/25/the-global-data-commons-gdc/> accessed 31 January 2020

[31] Narain (n 20)

[32] United Nations, 'High-Level Political Forum Goals in Focus – Goal 17: Strengthen the Means of Implementation and Revitalize the Global Partnership for Sustainable Development' (2018) <https://unstats.un.org/sdgs/report/2018/goal-17/> accessed 31 January 2020

manage such impact, making it more feasible for policymakers and governance models to keep apace with technological change.

These new structures – the AI Commons, the Global Data Access Framework and the AI4SDG Centers – can help the international community reconcile technical, ethical, commercial, legal and operational frameworks and protocols – to take power of AI technologies and successfully make unified progress towards the achievement of the SDGs.

The Use of AI by Online Intermediation Platforms

Conciliating Economic Efficiency and Ethical Issues

*Frédéric Marty and Thierry Warin**

This paper focuses on the effects of the implementation of artificial intelligence-based algorithms by online intermediation platforms in terms of both economic efficiency and fairness or ethical dimensions. It addresses three main issues: the consumer segmentation and the capacity to discriminate; the strategic use of artificial intelligence by dominant platforms in co-opetitive digital ecosystems; and the role of artificial intelligence-based tools to guarantee trustworthy user-reviews on e-commerce platforms. This paper emphasises the importance of having strong guarantees for platform users in terms of transparency and accountability.

I. Introduction

Although still in its infancy, the academic literature in the economics discipline insists on the promises of Artificial Intelligence (hereafter AI).[1] Our focus here lies in the field of Industrial Organization. In the case of electronic platforms, an increasing use of AI can substantially improve performance in several areas. For instance, for search or matching purposes, AI can be used to refine the recommendations provided to Internet users, hence reducing the search costs as well as the coordination costs.[2] Another example is that AI can also contribute to improve the level of trust in the platform. Similarly, AI can provide the platform with advanced user dissatisfaction detection tools. AI can also help solve the 'cold start' problem for new platforms entering the market by providing incentives for consumers to give trustworthy reviews.[3]

We will also rely on the following definition of an online intermediation platform: a platform is constituted of users – consumers and suppliers – whose transactions are subject to direct and/or indirect network effects.[4]

However, this great power comes also with great challenges. AI could indeed exacerbate consumer manipulations or generate/worsen certain biases as predictions made about consumer preferences might have a self-fulfilling effect.[5] The high degree of consumer segmentation allowed by AI could also lead to discriminatory practices.[6] Such discrimination may take the form of price discrimination but also of differences in the quality of the products being offered.[7] Similarly, AI-based advanced indicators – set up by the platform to measure the quality of the service provided by its complementors – can be manipulated to impose unbalanced contractual terms. Finally, the role of AI in encouraging consumer opinions may

DOI: 10.21552/delphi/2019/4/11

[*] Frédéric Marty, CNRS – GREDEG – Université Côte d'Azur; CIRANO, Montréal. For correspondence: <frederic.marty@gredeg.cnrs.fr> Thierry Warin, SKEMA Business School; CIRANO, Montréal. For correspondence: <thierry.warin@skema.edu>

[1] Catherine Tucker, 'Privacy, Algorithms, and Artificial Intelligence' in Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds) *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press 2018) 423–37

[2] Thierry Warin and Daniel Leiter, 'Homogenous Goods Markets: An Empirical Study of Price Dispersion on the Internet' (2012) 4 International Journal of Economics and Business Research 514–29

[3] Paul Milgrom and Steven Tadelis, 'How Artificial Intelligence and Machine Learning Can Impact Market Design' (2018) NBER Working Paper n°24282

[4] Broekhuizen et al, 'Digital Platform Openness: Drivers, Dimensions and Outcomes'(July 2019) Journal of Business Research; Jean Rochet and Jean Tirole,'Platform Competition in Two-Sided Markets' (2003) 1 Journal of the European Economic Association 990–1029

[5] Pelle Guldborg Hansen and Andreas Maaløe Jespersen, 'Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy' (2013) 4 European Journal of Risk Regulation 3–28

[6] Marc Bourreau and Alexandre de Streel, 'The Regulation of Personalised Pricing in the Digital Era' SSRN Scholarly Paper ID 3312158, Social Science Research Network <https://papers.ssrn.com/abstract=3312158> accessed 20 January 2020

[7] Preston McAfee 2007, 'Pricing Damaged Goods' (2007)1 Economics: The Open-Access, Open-Assessment E-Journal.

have a dark side if used to generate fake reviews. Thus, the effects of using AI in online platforms may be more debatable than initially expected.

Therefore, the purpose of this contribution is to propose a framework to study the opportunities and risks associated with the use of AI in the specific context of online intermediation platforms. In particular, we highlight three priority areas that we consider of the utmost importance in terms of risks posed to the nature of competition. These priority areas are key since they are actual threats to a healthy market system. Indeed, while lawmakers have considered some of the ethical issues related to AI, attention must also be paid to its consequences on competition. AI can affect the competitive dynamics of markets by increasing the market power of dominant firms at the expense of their competitors, trading partners, and consumers.[8] While the use of AI can address some of the key competitive issues in the platform economy, such as cold start and contestability, it may also increase imbalances between operators and facilitate manipulation. These issues are not only linked to efficiency but also to the dynamics of competition itself. A dominant operator (or a gatekeeper) might control competitive forces and impair consumers' freedom of choice and competitors' access to market.

For instance, Hemphill (2019) underlines that the incumbent's advantages on digital markets can be exacerbated by the development of machine learning. Dominant platforms benefit from economies of scale and scope and from an access to users' data that the new entrant cannot easily replicate or overcome. Such potential barriers to entry may significantly increased by AI implementation. Such technologies may exacerbate the incumbents' competitive advantages because of their high fixed costs, their data-based performance nature and the huge investments realised by both internal and external growth (mergers and acquisitions).

After presenting the current state of regulatory conversations, our conclusion is that the risks posed to a healthy market system are at best underestimated, at worst totally ignored. In what follows, first, we analyse the proposals and regulations made by public authorities aiming at guaranteeing a trustworthy AI and second, we propose three examples of AI-based algorithms to illustrate ethical and economic trade-offs. If the invisible hand enters an AI black box, it is necessary to balance efficiency gains and risks for sound competition and economic freedom.

II. AI Accountability in the Face of Economic and Ethical Risks

To the best of our knowledge, nowhere in the regulations do we find a reference to the potentially disruptive impact of AI on the market economy. According to the French National Convention in 1793: 'great responsibility follows inseparably from great power'. Such a requirement is still of the utmost importance in this AI age. What is known as the 'AI revolution' is an incredibly powerful tool but it also poses a lot of questions. For the first time in our history, autonomous systems can perform complex tasks equivalent to natural intelligence. The term Artificial Intelligence was coined by John McCarthy in 1956 in a proposal for a summer research project to be held in Dartmouth in 1957.[9] AI constitutes a major form of scientific and technological progress, which can generate considerable social as well as economic benefits.[10]

AI can be understood as a general-purpose technology.[11] We propose to use the following updated definition of AI from the European Union High-Level Expert Group on AI:

> 'Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can adapt their behavior by analysing how the environment is affected by their previous actions. As a scientific discipline,

[8] Sypros Makridakis, 'The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms' (2017) 90 Futures 46–60 <https://doi.org/10.1016/j.futures.2017.03.006> accessed 30 January 2020

[9] J. McCarthy et al, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence' (1956) 13 http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf accessed 30 January 2020

[10] Ajay Agrawal, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Review Press 2018)

[11] Erik Brynjolfsson, Daniel Rock and Chad Syverson, 'Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics' in Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds) *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press 2018) 23–57

> AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)'.[12]

The development of AI does pose major ethical challenges and social risks. Indeed, AI can restrict the choices of individuals and groups, disrupt the organisation of labor and the job market, or influence politics to name but a few. Scientific progress brings incredible benefits while carrying new risks. Citizens must determine the moral and political ends that give meaning to the risks encountered in an uncertain, and complex world.

On 8 April 2019, the High-Level Expert Group on AI presented their Ethics Guidelines for Trustworthy Artificial Intelligence. According to the guidelines, the main principles for a trustworthy AI are threefold: (1) lawful (abiding by all applicable laws and regulations), (2) ethical eg respecting ethical principles and values, and (3) robust both from a technical and a social perspective.[13] In its 'EU guidelines on ethics in artificial intelligence: Context and implementation' report, the European Commission's High-Level Expert Group on AI proposes 7 key guidelines for AI systems should meet in order to be deemed trustworthy: Human agency and oversight, Technical Robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental well-being, and Accountability.[14] Such concerns resonate with the principles laid down by the Montreal Declaration (2018)[15]: Well-being, respect for autonomy, privacy and intimacy, solidarity, democratic participation, equity, diversity inclusion, prudence, responsibility, sustainable development.

The question of healthy markets relies on data governance as much as on sustainable development, two topics in HELG and the Declaration of Montreal. Data governance is also to be found in article 25 of GDPR[16], along the UNESCO's 7th principle within the Beijing Consensus: the 'impact of AI on people and society should be monitored and evaluated throughout the value chain'[17].

The focus on the guarantees related to the use of AI is also reflected by the OECD's Recommendation of the Council on Artificial Intelligence.[18] Its first section sets up the 'Principles for responsible stewardship of trustworthy AI'. The principles advocated by the OECD are based on a review of potential risks for the society while enabling an AI digital ecosystem.

Although all these points are relevant and of the utmost importance, these reports give little attention to the specific competitive risks associated to digital oligopolies. There is no mention about the organisation of the market economy in this AI age. However, drawing from the OECD AI Principles, the G20 adopted human-centered AI Principles in Japan in June 2019. This is maybe the only text that gets as close as we can imagine to the notion of AI being disruptive for the market economy.

If we focus on intermediation platforms, the risks associated with the use of AI can be considered from the perspective of competition or consumer protection laws, as well as the protection of personal data. Even if these legal resources can address a significant part of the risks, they cannot answer all the ethical issues raised. It will therefore be up to the lawmakers to strike a balance between potential efficiency gains and derived risks. We illustrate these trade-offs in the next three sections through examples of AI implementation by online platforms. We insist for each of them on the potential efficiency gains that can stem from AI but also on the ethical risks raised.

[12] HLEG-AI, European Commission, 'Ethics Guidelines for Trustworthy AI' (8 April 2019) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> accessed 20 January 2020

[13] ibid

[14] Tambiama Madienga, 'EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation – Think Tank' (2019) <http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163> accessed 20 January 2020

[15] Montreal Declaration, 'The Declaration – Montreal Responsible AI' (2018) <https://www.montrealdeclaration-responsibleai.com/the-declaration> accessed 20 January 2020

[16] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance) OJ L. Vol. 119. http://data.europa.eu/eli/reg/2016/679/oj/eng accessed 30 January 2020

[17] UNESCO, 'Beijing Consensus on Artificial Intelligence and Education - UNESCO Digital Library' (2019) <https://unesdoc.unesco.org/ark:/48223/pf0000368303> accessed 30 January 2020

[18] OECD, 'OECD Legal Instruments' (2019) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> accessed 20 January 2020

III. Using AI for a Finer Consumer Segmentation: Efficiency at the Risk of Discrimination?

The performance of an intermediation platform or a search engine is based on the ability of its algorithms to deliver recommendations tailored to the needs of its users.[19] AI allows a more refined understanding of the latter by attaching each user to a narrowly defined segment. At the very least, the proposed result is dedicated to each user based on the prediction that is made about her expectations or her ability to pay. These are undoubtedly pro-efficiency effects. Search costs are substantially reduced for consumers. However, the use of AI in search engines and matching platforms may be accompanied by a rising risk for consumers: channeling too narrowly her choice towards the option for which the adequacy prediction is strongest. In this context, the use of AI may help perpetuate biases. For instance, in an experiment about job discriminations, job ads for higher-paid positions were displayed 6 times more to men than to women.[20]

In addition, risks of a self-fulfilling prophecy (and of undue restriction of consumer freedom of choice) must be taken into account. Is the development of AI in this area likely to lead to confirmation and reinforcement of social biases? The assignment of a consumer to a particular pattern certainly makes it possible to send her offers corresponding to her needs but also has a performative effect by locking her up in a restricted space of choice. The algorithm has the effect of closing options to her and thus constraining her future options. Insofar AI is only a predictive tool, we can arrive at the paradox in which the algorithm options would be verified ex-post simply because the very consequence of the prediction is the restriction of the span of possible choices. In the same vein, the AI originated proposal is all the more likely to be accepted since it reinforces the consumer's decision-making bias.

It should also be noted that the ability to statistically infer from the observed data the maximum amount that the consumer is willing to pay could lead to the implementation of personalised prices. Without going as far as perfect discrimination in which, the price would be equal for each consumer to her propensity to pay, AI can lead to a very efficient price segmentation.[21] The economic effects of price discrimination are ambiguous.[22] It is favorable in terms of total efficiency and can allow – through cross-subsidies – for certain consumers to access the product. However, discriminatory prices result in a transfer of welfare from the consumers to the platform.[23]

The algorithm can also play on the range of products offered to the Internet user. Depending on its anticipated degree of expertise or the needs assigned to her, the technical performance of the proposed product may vary. The discrimination through versioning can rely not on prices but performance or quality. For a same price, the characteristics of the product offered might differ from a user to another. At the extreme, in case of an on-demand production (in a 4.0 industry world), the product can be dedicated to a specific user. Such an approach may lead to deceptive commercial practices and harmful discrimination among consumers. We could easily imagine that naïve or captive ones might only access to products characterized by deteriorated performances as a result of machine learning methods applied to consumer segmentation.[24]

A burgeoning field of the economics literature investigates the possible exploitative abuses that could affect the more naïve consumers.[25] Firms increasingly have the capacity to discriminate among their consumers. The ever-increasing flow of information, the enhanced capacities to process the data collected and the obfuscation of on-line prices and offers may lead to a quasi-perfect discrimination exposing naïve consumer to pay unexpected charges[26] or to access to inferior quality goods or services.

[19] (n 3)

[20] Tom Simonite,'Study Suggests Google's Ad-Targeting System May Discriminate' (*MIT Technology Review, 6 July 2015)* <https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/> accessed 20 January 2020

[21] Salil Mehra, 'Antitrust and the Robo-Seller: Competition in the Time of Algorithms' (2016) 100 Minnesota Law Review 1323-1375

[22] Hal Varian, 'Artificial Intelligence and Industrial Organization' (2018) NBER Working Paper

[23] J.-P. Dubé and Sanjong Misra, 'Scalable Price Targeting' (2017) Working Paper University of Chicago

[24] Niladri Syam and Arun Sharma, 'Waiting for a Sales Renaissance in the Fourth Industrial Revolution: Machine Learning and Artificial Intelligence in Sales Research and Practice' (2018) 69 Industrial Marketing Management 135–46.

[25] Paul Heidhues and Botond Kőszegi, 'Naïveté-Based Discrimination' (2017) 132 Quarterly Journal of Economics 1019-1054

[26] Xavier Gabaix and David Laibson, 'Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets' (2006) 121 Quarterly Journal of Economics 505-540

Perverse effects resulting from algorithm-based discrimination can also stem from amplifications of social biases. Actually, matching or price algorithms may confirm or aggravate some discrimination already existing in society. Two difficulties can be considered. The first difficulty echoes a situation in which the algorithm is based on reinforcing self-directed learning. As such, it learns from available data and evolves through interactions. In doing so, it risks reproducing social biases and, much worse, amplifying them. The second difficulty is related to the worsening of the economic consequences of discriminations. Many studies emphasise this impact on the income of agents offering their services on platforms or on the opportunities available to them.[27]

Consumers may react negatively to personalised prices prices or price strategies generating random prices.[28] For instance, Amazon had to abandon random price variation initiatives in 2000.[29] Thus, the reputational damage can be significant if the consumer has a perception of the platform's behavior as manipulative, misleading or unfair. In that case, the potential 'market-based' sanction can play as a price signal incentivising the platform to monitor carefully its practices. However, such a self-disciplinary effect can only be effective if the market position of the platform remains contestable (if the competition is still *one click away*) and if the consumer is effectively aware of these practices.

Consumers may adversely react to new marketing practices based on AI as the possible shift from a shopping-then-shipping model to a shipping-to-shopping model.[30] As retailers can more and more precisely predict consumers' future needs, they can send the product before any formal order, allowing the consumer to freely return the item.[31] However, an undesired shipping, not mandatory resulting from a false prediction about a consumer's needs and preferences, may lead to a negative reaction. For instance, a consumer's current preference may differ from the ones inferred from her past behavior. The proposed product may put her in an uncomfortable situation.[32] The consumer may also react negatively to a perceived loss in autonomy in terms of consumption choices.[33]

One of the main concerns raised by AI implementation for marketing recommendations in terms of consumers' reactions can be illustrated with the 'privacy-personalisation paradox'.[34] On the one hand, consumers ask for dedicated offers but, on the other hand, they want to preserve their privacy. Their tradeoff might be distorted considering imperfect rationality. According to Acquisti (2004)[35], consumers may be 'privacy myopic'. In other words, they may divulge a substantial amount of information in return for a not so substantial counterpart.[36]

The platform monitoring by third parties can be helpful. Distributed surveillance schemes can be a way to provide guarantees to consumers and to provide the proper incentives to promote a competition through quality (or through commitments on fair practices) among platforms.

Symmetrically, market-based incentives might be insufficient to guarantee that firms properly control the effects of their algorithmic-based decisions in terms of fairness. Despite their own bias, consumers may react negatively as soon as they perceive the

[27] Benjamin Edelman, Michael Luca and Dan Svirsky, 'Racial Discrimination in the Sharing Economy: Evidence from a field experiment' (2017) 9 American Economic Journal – Applied Economics 1-22;
Alex Rosenblat et al, 'Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination' (2017) 9 Policy and Internet 256-279;
Mingming Cheng and Carmel Foley, 'The Sharing Economy and Digital Discrimination: The Case of Airbnb' (2018) International Journal of Hospitality Management 95-98;
Grazia Cecere et al, 'STEM and Teens: An Algorithmic Bias on Social Media' (2018) Working Paper SSRN n° 3176168

[28] Thierry Warin and Daniel Leiter,'Homogenous Goods Markets: An Empirical Study of Price Dispersion on the Internet' (2012) 4 International Journal of Economics and Business Research 514–29; Akiva Miller, 'What Do We Worry About When We Sorry About Price Discrimination? The Law and Ethics of Using Personal Information for Pricing' (2014) 19 Journal of Technology Law and Policy 41-104

[29] Matthew Edwards, 'Price and Prejudice: The Case against Consumer Equality in the Information Age' (2006) 10 Lewis and Clark Law Review 559

[30] Thomas Davenport et al, 'How Artificial Intelligence Will Change the Future of Marketing' (2020) 48 Journal of the Academy of Marketing Science 24-42

[31] (n 10)

[32] (n 30)

[33] Quentin André et al, 'Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data' (2018) 5 Consumers Needs and Solutions 28-37

[34] Elizabeth Aguire et al, 'Unravelling the Personalization Paradox: The Effect of Information Collection and Trust-building Strategies in Online Advertisement Effectiveness' (2015) 9 Journal of Retailing 34-49

[35] Alessandro Acquisti, 'Privacy in Electronic Commerce and the Economics of Immediate Gratification' (2004) 5th Conference on Electronic Commerce, New York

[36] Joshua Gerlick and Stephan Liozu, 'Ethical and Legal Considerations of Artificial Intelligence and Algorithmic Decision-making in Personalized Pricing' (2020) Journal of Revenue and Pricing Management.

firm's behavior as unfair or manipulative. The perceived (un)fairness echoes with distributive justice-related concerns and may lead to sale losses.[37]

The use of AI-based recommendations systems raises reputation-related concerns for firms as well as ethical and legal ones. If an algorithm makes its predictions as a non-accountable black box, its developers and the firm using it can be liable about its potential discriminatory or unfair effects. Martin (2019)[38] explains that a firm may be accountable if it develops or implements an inscrutable algorithm. She defines such an algorithm as one that limits – or excludes – any human intervention in the decision process and makes this algorithm impossible to objectively explain ex post. In this framework, a firm engages its corporate responsibility by relying on 'too difficult to explain' decision processes. In other words, the lack of accountability may be seen as the opposite of an ethical-by-design approach. As a consequence, opaque and non-accountable algorithms require a supervision, from a third-party or a regulatory agency.[39] The confidence towards AI-based decisions may be reinforced by the recourse to XAI, eg explainable AI.[40]

We can also insist on a second potential advantage of AI in the field of search engines. AI-based search recommendations can be customised according to the person's natural search process (broad first and progressive refinement). Indeed, consumers' search behaviour evolves during its successive stages. AI-based algorithms can adjust their results to fit with such a process. Based on an analysis performed on e-Bay users, Blake and al (2015)[41] show that the consumers prefer at the first steps very broad-range results to screen the available options and progressively converge toward more narrowly targeted results. A search or matching algorithm may reproduce such path and revise at each iteration the scope and the characteristics of the results displayed. However, such a tool has two sides; it can both support and distort the consumer's choices. Again, ethical concerns must be addressed.

IV. Using AI to Facilitate Trust in Transactions: Correcting Information Asymmetries or Increasing Trading-Partners' Vulnerability?

Trust in online platforms is a complex matter. It depends on the nature of the platform (proprietary vs open-source), the industry, the technology (blockchain vs https), the company, etc.[42]

One of the digital intermediation platforms' key factors of success has been their ability to secure transactions. This trust is not only about securing payments but also in terms of reducing informational imperfections that could prevent the act of purchase. For consumers, these imperfections were due to incomplete and asymmetric information about the quality of products and sellers active in online marketplaces. The opinions submitted online have played a significant role in correcting these informational biases. AI can be an interesting relay to address this issue: for instance, by predicting the quality of a given seller by interpreting online exchanges written in natural language in previous transactions. The analysis of exchanges between customers and independent sellers (through natural language processing) can make it possible to construct advanced indicators of underperformance and therefore to intervene to remedy it very early on. As such, the platform protects its consumers evolving in incomplete and asymmetric information conditions.

Putting in place mechanisms to create an environment that would increase participants' confidence is one of the keys of the platforms' business model.[43]

[37] Timothy Richards, Jura Liaukonyte and Nadia Stretskaya, 'Personalized Pricing and Price Fairness' (2016) 44 International Journal of Industrial Organization 138-153

[38] Kirsten Martin, 'Ethical Implications and Accountability of Algorithms' (2019) 160 Journal of Business Ethics 835-850

[39] Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (Harvard University Press 2015)

[40] (n 36)

[41] Thomas Blake, Chris Nosko and Steven Tadelis, 'Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment' (2015) 83 Econometrica 155-174

[42] Carin van der Cruijsen, Maurice Doll and Frank van Hoenselaar, 'Trust in Other People and the Usage of Peer Platform Markets' (2019) 166 Journal of Economic Behavior and Organization 751–66 <https://doi.org/10.1016/j.jebo.2019.08.021> accessed 30 January 2020;
Imene Ben Yahia, Nasser Al-Neama and Laoucine Kerbache, 'Investigating the Drivers for Social Commerce in Social Media Platforms: Importance of Trust, Social Support and the Platform Perceived Usage' (2018) 41 Journal of Retailing and Consumer Services 11–19 <https://doi.org/10.1016/j.jretconser.2017.10.021 > accessed 30 January 2020;
Nuan Luo et al, 'Integrating Community and E-Commerce to Build a Trusted Online Second-Hand Platform: Based on the Perspective of Social Capital' (2020) 153 Technological Forecasting and Social Change 119913 <https://doi.org/10.1016/j.techfore .2020.119913> accessed 30 January 2020

[43] Li Yfan Macinnes Ian and Yurcik William, 'Reputation and Dispute in eBay Transactions' (2005) 10 International Journal of Electronic Commerce 27-54

Such a characteristic was not a foregone conclusion since trust could not be based on interpersonal relationships, the experience of past transactions or the collective control exercised by a given community or corporation. Nor could trust come from technical devices as is the case, for example, through blockchain technologies in which cryptographic evidence can replace trust. The success of the first online marketplaces was ensured by the implementation of buyer feedbacks. These opinions made it possible to benefit from an ex-ante and not only an ex-post evaluation of the quality of the products. In other words, the sharing of information meant that the goods and services consumed did not fall, for each successive consumer, within the category of experience goods. Even before entering into a transaction with a formerly untried merchant, buyers on a platform benefit from the assessments of the seller's previous consumers.

However, this success is the subject of increasing challenges. First, consumers have no incentives to spend time to write reviews and can individually behave as free riders. Second, their confidence in consumers' review available online tends to be decreasing. Such mistrust is due to biases in assessments, rating inflation[44] and risks of opinion manipulation[45]. Some users may also collude to artificially increase the ratings by relying on puppet consumers posting false opinions corresponding to false transactions.[46]

The use of artificial intelligence can be a lever to restore this trust.[47] The idea is to use the messages exchanged on the platform between sellers and buyers before and after the transaction. The support of NLP (Natural Language Processing) allows such an evaluation. The algorithm aims at predicting which characteristics each consumer is likely to appreciate in the light of her interests and needs.

Such a method was implemented by Masterov et al (2015)[48] on comments left on the eBay platform. It is a matter of finding an element that can predict an unsatisfactory transaction. The authors relied on messages and internal data within the platform that could indicate that the transaction was not satisfactory (complaint, non-receipt or return of the object). The indicator of bad experience is the dependent variable. The algorithm will aim to predict this result from the messages exchanged. After a transaction, no messages can be exchanged, negative messages can be listed, and finally 'neutral' messages can be recorded. 85% of the transactions studied do not generate any messages. When there is no message, the number of unsatisfactory transactions is 4%. When a neutral message is sent, this rate is 13%. When at least one negative message is sent, this rate increases to 30%. A priori, the higher the proportion of negative messages a seller receives, the less quality he can be considered as. This frequency makes it possible to calculate a quality score that appears to be a good predictor of future performance.

AI allows this indicator to be inferred from large databases of email exchanges written in natural language to prevent the consumer – disappointed by a seller – from turning away from the platform.[49] In the present case, 'the fraction of a seller's message traffic that was negative predicts whether a buyer who transacts with this seller will stop purchasing on eBay'. This ultimately allows the platform to sanction a non-performing seller on objective grounds or to have leading indicators of the deterioration in the quality of the service provided. These monitoring methods may also raise concerns as soon as we consider information and power asymmetries between the platform and its complementors. Although these tools ultimately protect consumers, they place independent sellers under even closer control of the platform. By doing so AI increases their dependence and vulnerability. What is good for consumers is not always good for platforms' trading partners.

Therefore, there are also ethical considerations with some algorithms punishing non-performing sellers. It is even more relevant in the context of the potential black-box effect, but solutions exist without opening the black box.[50] Explainable models can

[44] Georgios Zervas, Davide Proserpio and John, Byers, 'A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average' (2015) Working Paper Boston University

[45] Dina Mayzlin, Yaniv Dover and Judith Chevalier, 'Promotional Reviews: An Empirical Investigation of Online Review Manipulation' (2014) 104 American Economic Review 2421-2455

[46] Weijia You et al, 'Reputation Inflation Detection in a Chinese C2C Market' (2011) 10 Electronic Commerce Research and Applications 510-519

[47] (n 3)

[48] Dimitriy Masterov, Uwe Mayer and Stevn Tadelis, 'Canary in the E-commerce Coal Mine: Detecting and Predicting Poor Experiences Using Buyer-to-sellerMmessages' (2015) Proceedings of the 16th ACM Conference on Economics and Computation 81-93

[49] Chris Nosko and Steven Tadelis,'The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment' (2015) NBER working paper n°20830

[50] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2017) ArXiv:1711.00399 [Cs] <http://arxiv.org/abs/1711.00399> accessed 20 January 2020

be detected are often proposed as a solution to the potential black box issue.[51] Algorithms have to be accountable without opening the black box, mainly for competitive reasons related to trade secrecy.[52]

V. Using AI to Create a Market for Online Evaluations: In Search of Objectivity

Creating or reinforcing the trust granted to an online intermediary implies providing consumers with a large number of reviews on its products. Such opinions are essential to reduce consumers' informational asymmetries and by doing so addressing the cold start issue for a new entrant, for instance, an independent seller proposing its items for the first time on a platform. This last one cannot easily transfer its reputation from a platform to another because of the barriers to data or review portability. A new entrant has strong incentives to reward its users to write reviews; in this context, the question is to know how to conciliate these incentives with guarantees in terms of objectivity. AI may be used as a tool allowing the provision of unbiased incentives for online notifications.

AI may address the issue of the cold start of platforms or sites publishing editorial content online. In the case of marketplaces, comments are needed to create trust, but these comments must be trustworthy! Bad comments can be fake ones by which companies punish their competitors.[53] Good comments, on the other side, can be paid to specialised companies in writing false consumer reviews. AI may be used to create a market for online assessments. A large majority of buyers on online marketplaces leave no opinion. In absolute terms, the consumer has no reason to do so: it consumes time and for future purchases, she can adopt a stowaway strategy using the opinions of others. The problem is not just about individual incentives. As Milgrom and Tadelis (2018)[54] note, this is also an industrial organization problem: the low number of third-party evaluations on a new platform makes buying from it less secure than buying from a platform with a large 'stock' of opinions.

The European Commission's June 2019 Regulation on relationships between sellers and platforms stressed this point: the lack of data and review portability does not allow the seller to transfer her reputation from a marketplace to another.[55] Hindering consumers' reviews portability has two potential anti-competitive effects. First, it increases the seller's dependence on the platform (by increasing switching costs). Secondly, it constitutes a barrier to entry for new platforms (which also has the indirect consequence of depriving sellers of exit options compared to existing platforms and thus further increasing their dependence).

How to use AI to solve this problem? Li et al. (2016)[56] analyse from a case study on the Chinese Taobao platform the possibility to charge merchants for the option of having buyers leave a notice. The idea is not to buy good reviews. The problem is always one of trust. It is a question of entrusting an algorithm, and not the seller himself, with the task of deciding whether the opinion is relevant. The experiment began in March 2012 with an 'evaluation discount' scheme (taking the form of ex-post reimbursements or discount coupons). Payment is made whether the opinion is positive or negative. It is only the informational quality of its content that is taken into account. The interest is twofold. First, it makes it possible to distinguish between good and bad salesmen. Indeed, the purchase of appraisals is an investment that will only be profitable if and only if the seller is of good quality. As the seller knows her own type, this mechanism acts as a revealing contract. Second, it allows the seller to solve the problem of the cold-start issue. Investment in reputation can be accelerated by purchasing 'objective' assessments.

The same cold-start issue applies for online news publishers. As Yang et al (2019)[57] stress, the value of

[51] (n 39)

[52] Julia Dressel and Hany Farid,'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4 Science Advances eaao5580 <https://doi.org/10.1126/sciadv.aao5580> accessed 20 January 2020

[53] Justin Johnson and D. Daniel Sokol,'Understanding AI Collusion and Compliance' forthcoming in D. Daniel Sokol and Benjamin van Rooi (eds) *Cambridge Handbook of Compliance* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3413882> accessed 30 January 2020

[54] (n 3)

[55] European Commission, Regulation (EU) 2019/1150 on Promoting Fairness and Transparency for Business Users of Online Intermediation Services

[56] Lingfang Li, Steven Tadelis and Xiaolan Zhou, 'Buying Reputation as a Signal of Quality: Evidence from an Online Market Place' (2016) NBER Working Paper n° 22584

[57] Ze Yang et al, 'Read, Attend and Comment: A Deep Architecture for Automatic News Comment Generation' (2019) arXiv.org > cs > arXiv:1909.11974

an article closely depends on the number and the quality of the comments it generates. Commentaries provide additional information to readers and improve their engagement on the website. Editors have strong incentives to encourage such comments and debates among users. However, the use of AI-based technology might raise ethical issues. Opinions can be written by AIs and use the reactions of Internet users to create artificial fixing points or even guide debates. Such automatic news commenting systems might also aim at generating neutral and reliable comments enhancing the readers' experience by using 'read-attend-procedures' based on machine reading comprehension (MRC) devices.[58] It remains true, however, that the ethical guarantees required by firms are essential in guaranteeing the model integrity.

VI. Conclusion

What are the possible principles for guarantees associated with AI in economics? Can transparency be required[59] and be sufficient to make market players accountable? Allowing third parties to access the code might conflict with trade secrecy rules and increase the risk to see its algorithms fooled. How, then, can ex-post accountability for choices be ensured? Accountability demands the identification of three elements: (1) the people involved, (2) the decision process and (3) the inputs used to form this decision.[60] The increasing role of AI in platform economics supposes to provide guarantees that efficiency gains for firms will not be paid by increased information asymmetries and manipulation capacities at the expense of consumers and trading partners. The use of AI by electronic platforms must not facilitate exclusionary or exploitative abuses. Enhancing discrimination possibilities would be also problematic in that, as we have seen, the effect of price discrimination on consumer welfare can be discussed. Such potential abuses may compromise the confidence in the digital economy. These are not the only competition risks pointed out by the academic literature. Concerns about bot-led tacit collusion equilibria are also stressed. Self-reinforcing machine learning might favour spontaneous convergence of competitors toward collusive prices without any explicit intent and information exchange devices.[61]

Moreover, some additional and even more problematic dimensions should be considered. The first one concerns the freedom of choice for consumers and for producers, the access to the market. AI can, to some extent, constrain and manipulate choices without the accountability of algorithms being obvious at this time. The second one is related to the issue of the online reputation monitoring. AI can be a powerful tool to assess the nature and behavior of a seller on a marketplace and incentive him to provide a good service (through scorings or threats of account suspension or suppression). This evaluation mechanism does not focus on only one of the sides of the platform. The consumer himself can be given a score in her digital journey. This use of artificial intelligence raises questions in an area that is not exclusively the responsibility of the market but of individual freedoms. It offers new resources for monitoring individual behavior far beyond the sphere of online market transactions.

[58] ibid

[59] (30)

[60] Stephen Kosack and Archon Fung, 'Does Transparency Improve Governance?' (2014) 17 Annual Review of Political Science 65–87.

[61] Emilio Calvano et al, 'Algorithmic Pricing: What Implications for Competition Policy?'(2019) 55 Review of Industrial Organization 155-171

Sustainable AI Safety?

Nadisha-Marie Aliman, Leon Kester, Peter Werkhoven and Soenke Ziesche[*]

In recent years, the need to address the multi-faceted issue of AI governance with safety-relevant, ethical and legal implications at an international level is becoming increasingly critical. Simultaneously, the international community is facing a wide array of global challenges for which the United Nations initiated an agenda with 17 ambitious Sustainable Developmental Goals (SDGs). In this article, we analyse potential synergies between methodologies to tackle both the AI governance challenge and the SDG challenge and work out novel constructive recommendations for an SDG-informed AI governance and an AI-assisted approach to the SDG endeavour. However, we also expound multiple open issues and contextual limitations that might play a role. Overall, our analysis suggests that while sustainable AI Safety cannot be guaranteed and the goals and values of the international community may change with time, AI governance could aim at a sustainable transdisciplinary scientific approach instantiated within a corrective socio-technological feedback-loop. Finally, we elaborate on the importance of the SDGs related to education and strong institutions for the realisation of this potentially robust AI governance strategy.

I. Synergies Between the Challenges of UN Sustainable Developmental Goals and AI Value Alignment

As Ziesche has proposed,[1] it might be highly valuable to identify synergies between the so-called AI value alignment problem and the Sustainable Developmental Goals (SDGs) challenge which have so far largely been treated separately despite a potential mutual benefit. Thereby, the AI safety relevant problem of AI value alignment represents a crucial subtask for AI governance and aims at identifying methods on how to implement AI systems acting in accordance with human values. This problem of societal relevance has been acknowledged to be of highly complex nature due to the absence of sufficiently specific as well as universal human goals.[2] Complementarily, the SDGs could be for instance interpreted as representing a type of condensed compendium of certain key human values shared internationally across 193 states and thus offering a basis for AI governance. In addition, sufficiently value-aligned AI systems could be utilised as support to achieve the SDGs in a targeted way including support in policy making. In fact, these bidirectional synergies could be vital given the urgency to address AI governance issues and since the SDGs have been adopted in 2015 by the UN General Assembly in order to 'stimulate action over the next 15 years in areas of critical importance for humanity and the planet'.[3]

However, the UN SDG framework, which states that 17 SDGs should be achieved by 2030, reveals certain caveats that need to be considered a priori in order to be able to harness it for AI value alignment or to design AI systems directly supporting the framework. The 17 SDGs, including those related to poverty, environmental pollution or inequality are further

DOI: 10.21552/delphi/2019/4/12

[*] Nadisha-Marie Aliman, M.Sc., PhD candidate at Utrecht Univesity, Department of Information and Computing Sciences. For correspondence: <nadishamarie.aliman@gmail.com>;
Dr. Leon Kester, Senior Research Scientist on ethical intelligent systems, TNO Netherlands;
Prof. Dr. Peter Werkhoven, Professor at Utrecht University and CSO of TNO Netherlands;
Dr. Soenke Ziesche, Independent Researcher, India

[1] Soenke Ziesche, 'Potential Synergies Between The United Nations Sustainable Development Goals And The Value Loading Problem In Artificial Intelligence' (2018) Maldives National Journal of Research 47-56

[2] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014)

[3] UN General Assembly, 'Resolution Adopted by the General Assembly on 25 September 2015; 70/1: Transforming our World: The 2030 Agenda for Sustainable Development' (2015)

subdivided into 169 targets whose achievement is monitored via 232 indicators with varying quality. The differences in quality are partly reflected in the subdivision of the indicators into three different tiers. As of 26 September 2019, countries do not regularly produce data for 89 (so-called tier II indicators) out of the 232 indicators, while no internationally established methodology is yet available for a further 33 indicators (so-called tier III indicators).[45] One of the main issues is that several targets are not quantified and to specify indicators for such targets is particularly challenging. Despite these notable challenges, we propose considering the UN SDGs as complementary approach towards the AI Value Alignment problem. In order to achieve that, the set of SDGs has to be formulated in a machine understandable version to facilitate goal-oriented AI-based solutions. In order to identify for AI value alignment purposes what a society wants (ethical self-assessment) and in a second step what a society should want (ethical debiasing), it has been suggested to combine a scientifically grounded assessment of human ethics with technological methods such as virtual reality studies for experiences from a first-person perspective.[6] Thereby, we believe that the SDGs could serve as a heuristic able to supplement ethical self-assessment by qualitatively specifying candidate human values. Moreover, certain more precise SDG indicators might provide helpful quantitative targets in some cases. Beyond that, we will also discuss how the SDGs related to strong institutions and quality education are expedient for a robust dynamic approach to AI governance which is not only proactive but also foresees the need for reactive corrections leading to a socio-technological feedback-loop.[7]

In Section II we discuss possible contributions of SDGs for AI value alignment by taking the example of value alignment for intelligent autonomous systems and more precisely the autonomous vehicle case for illustrative purposes. In Section III, we comment on limitations and emerging sustainability challenges in this context and formulate a set of recommendations which also encompasses the other direction of the synergy, namely AI systems for UN SDGs. Finally, in Section IV, we conclude and discuss future prospects. In a nutshell, we do not claim that the SDGs are a comprehensive solution for AI governance, but rather a promising complementary tool given the urgency of the problem as well as the fact that the SDGs can be seen as the most detailed as well as inclusive vision for human development ever compiled.[8]

II. Complementing Value Alignment for Intelligent Autonomous Systems with UN SDGs

After having theoretically motivated the potential usefulness of UN SDGs for AI value alignment, we discuss the application of this proposition in the context of intelligent autonomous systems utilising the use case of autonomous vehicles (AVs) as helpful toy model with ethical, legal and environmental dimensions pertaining to the realisation of the SDG endeavour itself.[9] (In the following, we will refer to intelligent autonomous systems with the expression 'artificial intelligent system' instead, since we want to stress that the goals for decision-making in this context are specified by humans and irrespective of the level of automation, it is not the artificial system that crafts its own goals autonomously as often mistakenly assumed.) We use value alignment with AVs as toy model due to the fact that the use case exhibits domain-general important safety-critical, ethical and legal features many of which would pertain to the value alignment of a wide range of artificial intelligent systems deployed in real-world environments. Firstly, it reveals the need to make human values explicit for risk assessment and planning which represents a societal challenge of ethical self-assessment since humans are often reluctant to clearly express what they want. Secondly, the use case points to an-

[4] United Nations, 'Tier Classification for Global SDG Indicators' (2019)

[5] An exemplary tier II indicator is 14.1.1 (*index of coastal eutrophication and floating plastic debris density*) while the indicator 12.4.2 (*hazardous waste generated per capita and proportion of hazardous waste treated, by type of treatment*) represents an example for a tier III indicator.

[6] Nadisha-Marie Aliman and Leon Kester, 'Extending Socio-Technological Reality for Ethics in Artificiall Intelligent Systems' (2019) IEEE AIVR

[7] Nadisha-Marie Aliman, Leon Kester, Peter Werkhoven and Roman Yampolskiy, *Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. In International Conference on Artificial General Intelligence* (Springer 2019) 22-31

[8] (n 1)

[9] Ricardo Vinuesa et al, 'The Role of Artificial Intelligence in Achieving the Sustainable Development Goals' (2019) arXiv preprint arXiv:1905.00501

other challenge of scientific nature which is to design suitable machine-readable frameworks that can serve as scaffolds and templates for the identified human ethical values and legal conceptions. Thirdly, it might necessitate a societal-level aggregation of heterogeneous and often conflicting views within this type of ethical frameworks. Fourthly, due to its complexity, it might require a cognitive-affective extension of society (eg using targeted virtual reality studies[10]) facilitating a high-quality ethical self-assessment and ethical debiasing which constitutes a scientific and technological challenge. Fifthly, while the case might seem to correspond to a rather narrow domain, it has implications that extend beyond it and will need a supportive context which can be characterised as an institutional, legal and societal challenge.

Since the UN SDGs themselves, as well as its targets, might be too abstract to identify how they can be *directly* applied to the AV case, it is helpful to scan the SDG indicators[11] in a bottom-up fashion. In the following, we only mention a non-comprehensive exemplary set of some of the most straightforward related indicators. Regarding environmental awareness for AVs, one can for instance identify the indicators 9.4.1 (CO2 emission per unit of value added) and 11.6.2 (annual mean levels of fine particulate matter *(*eg PM2.5 and PM10*)* in cities (population weighted)). These indicators might be relevant for hybrid-electric AVs but also electric AVs that obtain their energy from correspondingly polluting sources. At the top-level, the indicator 9.4.1 is related to the SDG 9 which seeks to 'build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation', while indicator 11.6.2 stems from the SDG 11 which aims to 'make cities and human settlements inclusive, safe, resilient and sustainable'. Concerning ethical and legal aspects, one can for instance name indicator 3.6.1 (death rate due to road traffic injuries), 5.1.1 (whether or not legal frameworks are in place to promote, enforce and monitor equality and non-discrimination on the basis of sex), 16.7.2 (proportion of population who believe decision-making is inclusive and responsive, by sex, age, disability and population group) as germane in this context. These indicators are related to SDG 3 which aims to 'ensure healthy lives and promote well-being for all at all ages', SDG 5 '*achieve gender equality and empower all women and girls*' and SDG 16 on peace, justice and strong institutions respectively. All mentioned indicators are tier I indicators (ie relatively clearly formulated, respecting international standards and with regular updates on data available) except for 5.1.1 and 16.7.2 which are tier II indicators. Independently of the specific type of ethical framework envisaged for meaningful control of AVs, the presented indicators related to 5 SDGs could be helpful even though certainly not in isolation. To explain how they could be harnessed for an ethical framework for AVs, we first describe a recently introduced scientifically grounded non-normative framework for ethics in artificial intelligent systems denoted augmented utilitarianism[12] before linking it back to the SDG-related synergy.

Recently, augmented utilitarianism has been proposed as scaffold and template to fill in human values and as instrument to control artificial intelligent systems in a novel utility-based manner. Augmented utilitarianism is in accordance with modern insights in constructionist accounts of moral psychology[13] and cognitive neuroscience[14] according to which mental states (also moral judgements)[15] are embodied constructions based on domain-general processes of context-sensitive, perceiver-dependent, time-dependent and affective nature.[16] For this purpose, augmented utilitarianism introduces a type of context-sensitive and perceiver-dependent utility function that extends beyond the classical consequentialist and utilitarian utility functions which are focused solely on the outcome of actions. In this way, it allows a coalescence of the classical normative ethical

[10] (n 6)

[11] United Nations Statistical Commission, 'Global Indicator Framework for the Sustainable Development Goals and Targets of the 2030 Agenda for Sustainable Development; UN Resolution A/RES/71/313' (2017) <https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework_A.RES.71.313%20Annex.pdf> accessed 20 January 2020

[12] Nadisha-Marie Aliman and Leon Kester, 'Requisite Variety in Ethical Utility Functions for AI Value Alignment' IJCAI AI Safety Workshop 2019

[13] Chelsea Schein and Kurt Gray, 'The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm' (2018) Personality and Social Psychology Review 32-70

[14] Ian R. Kleckner et al, 'Evidence for a Large-scale Brain System Supporting Allostasis and Interoception in Humans' (2017) Nature Human Behaviour 0069; Suzanne Oosterwijk et al, 'States of Mind: Emotions, Body Feelings, and Thoughts Share Distributed Neural Networks' (2012) NeuroImage 2110-2128

[15] (n 12)

[16] Lisa Feldman Barrett, *How Emotions are Made: The Secret Life of the Brain* (Houghton Mifflin Harcourt 2017)

views related to virtue ethics, deontology and consequentialism – which seem to all possibly play a role in human moral judgements.[17] To achieve this, augmented utilitarianism offers a perceiver-dependent template allowing the joint consideration of agent, action and patient. For a meaningful control of artificial intelligent systems using this framework, people would not need to agree on what they value and how they weigh what they value. The main necessary precondition would be to consent to an acceptable superset of parameters allowing an aggregation of the perceiver-dependent and context-sensitive utility functions respecting legal constraints. (Note that these machine-readable utility functions would facilitate interpretability of reasoning/planning at the level of the decision-making component via the transparent human-crafted formulation of parameters and weights enabling concrete counterfactual comparisons.[18] However, interpretability at the sensor-level remains an important outstanding challenge.) The necessary ethical self-assessment and ethical debiasing to craft these utility functions can be assisted by experts from the legislative and be supported by technology such as virtual or augmented reality[19] providing a rich counterfactual experiential testbed for a responsible human-centred decision-making. To make justice to the time-dependency of human ethical conceptions, one would also need to update these augmented utility functions. This indispensable correction of utility functions paired with the need to update the world models of the AI systems themselves instantiates a dynamic socio-technological feedback-loop.

However, it becomes clear that such a general mechanism of correction of error within a socio-technological feedback-loop which is highly relevant for AI value alignment cannot succeed if the mentioned SDG 16 related to peace, justice and strong institutions is not realised to a sufficient degree. This is agnostic of the ethical framework considered, since the fact that human knowledge is prone to errors makes a correction process mandatory. Therefore, one might categorise SDG 16 as a meta-goal for AI governance. Furthermore, the SDGs identified can also provide more detailed information related to concrete parameters specifically applied to the AV case. Since society would need to specify a superset of candidate parameters that are admitted for consideration, the SDG indicators specified can help to extend or filter this superset. For instance, it might be recommendable to add CO2 and fine particulate matter related parameters in the augmented utility functions of the AVs if suited (even if the provided indicators 9.4.1 and 11.6.2 are rather restricted with regard to all climate change relevant measures) which is in the spirit of sustainable mobility. An obvious additional important parameter is related to road traffic injuries as encoded in the SDG indicator 3.6.1. Finally, one must address risk assessment parameters which are necessary because collisions can in practice not be avoided with absolute certainty at any time[20] and there is no 100% secure system[21] even if AVs are meant to drastically improve the security of mobility. Obviously, the UN SDGs do not allow a direct consideration of this case since crafted for a fully different purpose, although more generally, the indicator 5.1.1. and 16.7.2 reflect recommendations on gender-inclusive legal enforcement and non-discriminatory decision-making. However, this indication does not directly solve the complex problem of identifying parameters that could be relevant for dilemmatic situations in the context of risk assessment, an important part of AI Value alignment. We apply a closer analysis to this missing piece of crucial importance in Section III. However, these indicators might emphasise the general necessity to competently address discrimination based on algorithmic biases which we will touch upon in Section III. Lastly, one drawback of the SDG framework is that it does not allow the identification of precise weights and the establishment of concrete priorities in the pursuit of the SDGs. In total, it can be summarised that the UN SDGs allow a powerful supplement to value alignment with AVs (and more generally artificial intelligent systems) which add important qualitative and quantitative contributions. However, it is not meant as a standalone solution and should be utilised in conjunction with an ethical framework able to model ethical and legal dimensions and be extended by scientifically grounded and technology-

[17] Veljko Dubljević, Sebastian Sattler and Eric Racine, 'Deciphering Moral Intuition: How Agents, Deeds, and Consequences Influence Moral Judgment'(2018) PloS one e0204631

[18] (n 7)

[19] (n 6)

[20] Sixian Li et al, 'Influencing Factors of Driving Decision-Making Under the Moral Dilemma' (2019) IEEE Access 104132-104142

[21] Roman V. Yampolskiy and M. S. Spellchecker, 'Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures' (2016) arXiv preprint arXiv:1610.07997

assisted ethical self-assessment and debiasing measures.

III. Sustainability Challenges in the Context of AI Value Alignment

It is highly important to address the mentioned point of decision-making under dilemmatic circumstances, since while we exemplarily refer to the AV case as toy model, the topic is generally relevant for artificial intelligent systems and artificial decision support systems in critical domains where the lives and the well-being of people are inherent part of the decision process. Conceivable relevant application areas may be eg justice, medicine and bureaucracy but could also pertain to future human-machine collaboration forms such as human-robot rescue teams, hybrid fire brigades or even advanced domestic robots. Coming back to the AV case, it is also noteworthy that failing to address this issue could have non-trivial repercussions on a few SDG indicators themselves. If the satisfaction of society with proposed ethical guidelines for AVs is low, it might (ceteris paribus) slow down the acceptance of the technology and people would be less willing to switch to AVs. In turn, this reservation could possibly hinder an optimal overall reduction of air pollution (related to SDG indicators 9.4.1 and 11.6.2) and importantly, it is thinkable that the number of deaths due to road traffic injuries (see SDG indicator 3.6.1) which AVs are supposed to decrease could therefore not be decreased optimally. In fact, according to a study analysing the social dilemma encountered with AVs,[22] while people would in theory approve AVs equipped with a utilitarian approach to dilemmatic scenarios, they would not like to ride such an AV themselves. Moreover, people expressed their unwillingness to accept regulations mandating a utilitarian self-sacrifice of AV passengers and expressed their aversion to buy AVs in the presence of such regulations. This type of mechanisms could lead to the mentioned undesirable repercussions on some SDG indicators. In the following, we portray why the utilitarian approach to ethical dilemmas in AVs as eg suggested by German ethical guidelines stating that in unavoidable accident scenarios personal features (eg age) should not be considered[23] poses additional problems of different nature. Thereafter, we provide a set of recommendations on how to address such socio-technological issues by initiating an active societal debate supported by science and technology including AI systems themselves – finally linking it to the other direction of the synergy of AIs for UN SDGs.

One can distinguish two main types of problems that can arise when adopting a purely utilitarian decision-making for AVs but also more generally for artificial intelligent systems in critical domains: the first one is related to the discrepancy between the (often culture-dependent)[24] ethical intuitions of most people and the utilitarian approach and the second one concerns a fundamental problem[25] related to impossibility theorems for classical utilitarian utility functions. First, multiple experiments assessing ethical dilemmas with AVs have been performed eg in text form or virtual reality environments. Depending on the type of constellation and the focus of different recent virtual reality-based experiments,[26] the moral judgements or moral actions of participants (denoted as perceivers in the following) were heterogeneous and partly contradictory overall. In these experiments elements that were decisive included for instance: the perceived nature and transparency of the agent, the legal liability of the agent, whether the accident happened by action or by inaction, whether the action involves a self-sacrifice, the number of patients, the age of patients, the personality traits of the perceiver, the culture of the perceiver and the amount of time the perceiver had for a decision[27]. This is not surprising, since moral judgments are related to a perceiver-dependent dyadic cognitive template encoding a continuum along which an intentional agent is perceived to cause harm to a vulnerable patient[28]. The more this seems to be the case, the more immoral does the act seem to the perceiver. Thereby, the vulnerability people ascribe to patients can vary ex-

[22] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, 'The Social Dilemma of Autonomous Vehicles' (2016) Science 1573-1576

[23] Noa Kallioinen et al, 'Moral Judgements on the Actions of Self-driving Cars and Human Drivers in Dilemma Situations from Different Perspectives' (2019) Frontiers in Psychology <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02415/full>> accessed 20 January 2020

[24] Edmond Awad et al, 'The Moral Machine Experiment' (2018) Nature 59-64

[25] Peter Eckersley, 'Impossibility and Uncertainty Theorems in AI Value Alignment (Or Why Your AGI Should not Have a Utility Function)' (2018) arXiv preprint arXiv: 1901.00064

[26] (n 6)

[27] (n 6)

[28] (n 13)

tremely. Generally speaking, the way people perceive the agent, the action and the patient can vary with regard to a plurality of parameters of eg cultural, social, temporal, psychological and affective nature. Therefore, while the number of victims in an unavoidable collision certainly is an important factor to consider in ethical guidelines, human ethical intuitions tend to encompass a richer set of information. Finally, it is important to note that classical consequentialist and utilitarian utility functions have been shown to represent a safety risk if used in critical domains with future human well-being and human lives as part of the decision-making if used without more ado.[29]

As introduced in Section II, augmented utilitarianism allows a context-sensitive and perceiver-dependent account of human ethical intuitions which is not affected by the limitations encountered by utilitarian utility functions. Thus, AI Value Alignment could profit from harnessing this framework in addition to the mentioned SDG indicators and initiate a societal-level debate on the choice of a suitable superset of values that matter in dilemmatic circumstances and how they need to be weighted. However, while this would serve to tackle value alignment at the level of the decision-making component, artificial intelligent systems also need to exhibit value-aligned properties at the sensor-level. In the AV case, this would map by way of example to the problem of discrimination via algorithmic biases at the level of image classification. Next to the mentioned SDG indicators 5.1.1, 16.7.2 on gender-inclusive legal enforcement and non-discriminatory decision-making, one could add the tier II indicator 16.b.1 (Proportion of population reporting having personally felt discriminated against or harassed in the previous 12 months on the basis of a ground of discrimination prohibited under international human rights law). While it is important to strive for datasets with a large variety to forestall such often unintentionally arising discriminations, we stress that this can and should be complemented by an *explicit* formulation within the algorithm itself. Due to the nature of human ethical intuitions, a utility function that does not encode affective and dyadic parameters of the *current* society cannot be a good model for an ethical framework and can thus not instantiate a value alignment effort.[30] In many cases, this can manifest itself by leading to input-to-output mappings that people categorise as discriminatory. An example for such discriminatory mappings is the case where the picture of persons whose phenotype was underrepresented in the dataset was labelled with the class 'gorilla' by Google Photos. Another example is a study which was related to the AV context in which researchers analysed multiple image recognition systems and found that the images of pedestrians with darker skin tones were detected with a lower accuracy.[31] Next to more diverse datasets, it is indispensable to eg explicitly weigh misclassifications errors of the algorithms affectively. Not all misclassifications are equally important. In simplified terms, it is easily conceivable that for humans it makes a difference whether an image recognition system misclassifies a chimpanzee image as a gorilla in comparison to the case of a human being mistaken for a gorilla. However, many algorithms nowadays are implemented agnostic to analogies of such nuances. (As 'solution' for the mentioned incident, Google Photos opted to censor the gorilla label[32] as well as a few related labels including 'chimpanzee'.) If machine learning systems or artificial intelligent systems optimise on loss functions, objective functions or utility functions devoid of relevant affective, contextual and societal factors, undesired discriminatory side effects could occur continuously. (Note that this analogously applies to rule-based systems and others.) This would represent negative repercussions on both AI Value Alignment and UN SDGs. Seen from a different angle, it can be said that research on discrimination stemming from algorithmic biases would unify the directions UN SDGs for AI value alignment and AI for UN SDGs. An additional important aspect to cover for this type of research are so-called ethical adversarial examples which represent adversarial attacks on AI systems attempting to entice AI systems 'to action(s) or output(s) that are perceived as violating human ethical intuitions.'[33]

As already described, the SDG framework unfortunately exhibits a lack of precision for multiple indicators. Furthermore, certain of them are underspec-

[29] (n 26)

[30] (n 12)

[31] Benjamin Wilson, Judy Hoffman and Jamie Morgenstern, 'Predictive Inequity in Object Detection' arXiv preprint arXiv:1902.11097 (2019)

[32] Tom Simonite, 'When it comes to Gorillas, Google Photos Remains Blind' Wired 1 November 2018 <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/> accessed 17 October 2019

[33] (n 12)

ified. This makes it difficult to track progress towards specific indicators and top-level SDGs. However, it has been postulated that machine learning applications could extend the SDG indicators by utilising multimodal data from diverse sources for a better assessment of progress.[34] This could also be relevant if one uses AI as decision-support for policy-making that should be in line with the SDGs. Moreover, a dedicated type of positive computing could target SDG 3 in a broader sense (ensure healthy lives and promote well-being for all at all ages)[35]. However, so far, not many systematic AI attempts towards the SDGs and their targets have been reported yet[36]. From the perspective of AI value alignment for artificial intelligent systems, the identification of precise criteria based on which one would in the first place select SDGs or SDG indicators given a generic domain is non-trivial, since the SDGs have been motivated and formulated from an international perspective. While for the AV toy model we heuristically scanned the indicators in a bottom-up fashion searching for obvious matches, future work could develop a more sophisticated methodology. For instance, an important SDG that might as first glance seem unrelated to value alignment in the AV case in particular or to artificial intelligent systems in general, is the SDG 4 (ensure inclusive and equitable quality education and promote lifelong learning opportunities for all). As one can already extract from the article so far, it is highly recommendable to apply a transdisciplinary methodology to both AI value alignment and to the SDG challenge to avoid blind spots and a negligent approach to future global challenges. In the following, we comment on the importance of SDG 4 for AI governance and finally link it to SDG 16 on peace, justice and strong institutions.

We think that education and life-long learning – eg transdisciplinary further education for AI Safety and AI researchers as well as for authorities involved in AI regulation, and education fostering an awareness of socio-technological challenges for the general public – are highly powerful tools for both challenges. First, it provides a basis for the generation of novel approaches to AI governance. In fact, while some people believe that the goal in AI governance should be to achieve a consensus, a broad variation of scientific approaches represents an ideal breeding ground for progress. Second, a proactive AI governance approach is not enough due to errors and changes in human values that will occur, which means that one cannot solely rely on current strategies. Thus, it will be convenient to accumulate broad knowledge that might be helpful in the face of novel unpredicted problems that arise. Any AI governance approach therefore needs to be updatable by design in order to allow a corrective socio-technological feedback-loop. Unfortunately, the SDG framework is not meant to be steadily updated which represents a clear limitation that should be thoroughly taken into consideration when attempting to achieve its fixed goals. For instance, new unforeseeable challenges may be related to developments in AI itself (and other new technologies) as can be seen when considering the current SDG target 8.5, which aims to 'achieve full and productive employment and decent work for all women and men' – which against the background of technological advances might be neither realistic nor worthwhile any more[37]. Third, an education of the general public might be important, since many people exhibit ethical biases based on incorrect assumptions. In the AV case, this could for instance include anthropomorphism, presumed level of intentionality and agency or misconceptions on the functioning of AVs.[38] These epistemic gaps can be addressed via a more in-depth education leading to a more informed experience and ethical debiasing which respects the manifestation of moral pluralism known from psychology.[39] Overall, we believe that a scientifically grounded approach to AI governance supplemented by education is absolutely necessary given future challenges. However, we want to re-emphasise that without strong institutions as captured in SDG 16 which we termed an important meta-goal for AI value alignment, the mentioned strategies would be highly limited in their field of action. On the other hand, failing to address AI governance could lead to AI Safety risks with negative repercus-

[34] Niheer Dasandi and Slava Jankin. Mikhaylov, 'AI for SDG-16 on Peace, Justice, and Strong Institutions: Tracking Progress and Assessing Impact'(2019) Position Paper for the IJCAI Workshop on Artificial Intelligence and United Nations Sustainable Development Goals

[35] (n 1)

[36] Soenke Ziesche, 'Innovative Big Data Approaches for Capturing and Analyzing Data to Monitor and Achieve the SDGs' (2017) Report of the United Nations Economic and Social Commission for Asia and the Pacific: Subregional Office for East and North-East Asia (ESCAP-ENEA)

[37] (n 1)

[38] (n 6)

[39] (n 13)

sions to the SDG framework ranging for instance from compromising human well-being to existential risks in some cases[40].

IV. Conclusion and Future Prospects

Overall, one can conclude that it is expedient to embrace the SDGs and their general intention as a complementary foundation for the AI Value Alignment problem, yet one needs to acknowledge given limitations including the need for a revised/special version of the indicators to become fit-for-purpose. Against the background of our analysis, one can establish that the SDG framework exhibits two main weaknesses when applied to the AI value alignment challenge. First, the SDGs do not mention artificial intelligence at all, neither its significant opportunities, nor its significant risks, although both were to an extent known at the time when the SDGs were formulated. One reason for this is that these discussions were siloed in academic circles, and only recently the (now even more urgent) need for AI Governance has been acknowledged[41]. Second, human challenges and values change over time and unforeseeable factors might emerge, while the SDGs have no mechanism for an amendment until 2030, which is only justified by pragmatic reasons. This can be also illustrated by the predecessor of the SDGs, the Millennium Development Goals, which had partly different ambitions. Importantly, the above issues are intertwined. For example, new unforeseeable challenges may as well be related to developments in AI itself and other new technologies.

As stated by Karl Popper, 'no society can predict, scientifically, its own future states of knowledge.'[42] Hence, AI Safety cannot be guaranteed to be sustainable in the long run nor will the goals pursued by the UN necessarily remain unchanged. Nevertheless, we believe that it is a sustainable transdisciplinary scientific approach that one should strive for in order to efficiently tackle AI Governance and exploit the described beneficial synergies with the SDGs. For security and safety, one needs requisite knowledge at the right time. For this reason, one can argue that the SDG 4 on quality education and life-long learning contains a key element. However, in the light of the above, it seems imperative to additionally aspire to a corrective socio-technological feedback-loop enabling both proactive and reactive measures and for which SDG 16 on strong institutions represents a precondition.

[40] Soenke Ziesche, AI & Global Governance: A Seat at the Negotiating Table for AI? Opportunities and Risks, United Nations University 2 August 2019 <https://cpr.unu.edu/a-seat-at-the-negotiating-table.html> accessed 17 October 2019

[41] Allan Dafoe, *AI Governance: A Research Agenda' Governance of AI Program, Future of Humanity Institute* (University of Oxford 2018)

[42] Karl R. Popper, *The Poverty of Historicism* (Routledge 1957)

ERA
Academy of European Law

Artificial Intelligence (AI) and Human Rights

Brussels, 12-13 March 2020

Key topics

- AI technology
- Issue of responsibility and transparency of AI systems
- European Commission Ethics Guidelines for Trustworthy AI
- Relevant ECtHR case law
- Council of Europe Guidelines on Artificial Intelligence and Data Protection
- Impact of AI on selected human rights

Event number: 420D77 • **Language:** English

Artificial Intelligence (AI) and the Criminal Justice System

London, 4-5 June 2020

Key topics

- AI and its implications for the criminal justice system
- Using AI in criminal intelligence
- AI to predict crimes
- The internet industry approach to AI
- AI, ethics and privacy

Event number: 320D61 • **Language:** English

Summer Course on European Information Technology Law

Trier, 14-18 September 2020

Key topics

- Internet regulation in the European Union
- Role, responsibility and accountability of online platforms
- Regulating e-commerce in Europe
- Information technologies and intellectual property law
- EU legal framework applicable to data: sui generis protection of databases, trade secrets protection, free flow of data, access to data, data privacy in the digital world, data concentration
- Online and smart contracts
- Private international law and information technologies
- Digital identity and electronic payments
- Cybersecurity

Visit to the European Court of Justice

Participants in this summer course will also have the opportunity to attend a hearing at the CJEU in Luxembourg.

Who should attend?

Practitioners seeking an introduction to European IT law

Event number: 220B17 • **Language:** English

UP GRADE YOUR LEGAL EXPERTISE

Contact
Academy of European Law
Philipp Gonder
Tel. +49 (0) 651 937 37-811
E-mail pgonder@era.int
www.era.int

For more information, please visit our website

www.era.int

Events Spring/Summer 2020

20 YEARS lexxion

February

■	27 – 28	Fundamentals of State Aid Law | Amsterdam
▨	27 – 28	Public Procurement Requirements for ESI Funds | Paris

March

■	03 – 04	State Aid Requirements for Services of General Economic Interest | Brussels
■	12 – 13	State Aid Procedures - The Legal Obligations and Rights of the Parties Involved | Lisbon
■	19 – 20	How to Most Effectively Use Technical Assistance for ESI Funds Now and in 2021-2027 | Florence
■	25 – 26	Indicators, Monitoring and Evaluation in ESIF Programme and Project Management | Milan
▨	25 – 27	Spring Course: Public Procurement from A-Z | Milan
■	26 – 27	Programme and Financial Management of ESI Funds 2021 – 2027 | Nice
■	26 – 27	Effective Programming and Implementation of Financial Instruments | Malaga
■	30.3.–1.4.	Spring Course: Simplified Cost Options for ESI Funds | Catania

April

■	01 – 03	Real-life On-the-spot Visits: How to Detect and Combat Irregularities & Fraud in ESI Funds | Maastricht
■	02 – 04	Verwaltung und Prüfung der Europäischen Struktur- und Investitions-Fonds (in German) | Hamburg
■	02 – 04	EStAL Seminar | Catania
■	22 – 24	Spring Course: Concept & Principles of State Aid | Athens
■	22 – 24	Advanced Spring Course: Innovative Approaches to Risk Analysis, Verifications and Audits in ESI Funds | Nice
■	23 – 24	Migration and Security Funds: How to Effectively Manage, Audit and Control AMIF & ISF Now and in 2021-2027 | Nice
▨	23 – 24	Essentials of e-Procurement | Rome
■	28 – 29	State Aid for Agriculture, Forestry, and Rural Areas | Brussels

May

■	12 – 13	State Aid for Research, Development and Innovation Projects | Brussels
■	14 – 15	How to Conduct Management Verifications Most Effectively | Bologna
■	27 – 29	Summer Course: Financial Control and Audit in ESI Funds Now and in 2021-2027 | Sicily
■	28 – 29	Risk Management and Anti-Fraud Game of Jo Kremers (Edition 2020) | Sicily NEW
▨	28 – 29	Essentials of Public Procurement | Florence

June

■	09 – 10	State Aid: Assessment and Evaluation | Brussels
■	09 – 10	How to Successfully Manage, Control and Audit IPA Funds | Amsterdam
■	11 – 12	ESI Funded Agriculture Programmes and Projects | Catania
■	11 – 12	EStALI Interactive Forum on EU State Aid Law | Brussels
■	17 – 19	Summer Course: Irregularities and Fraud in ESI Funds and Public Procurement | Athens
■	17 – 19	Summer Course: Effective Usage of EU Financial Instruments Now and in 2021-2027 | Lake Como
▨	18 – 19	Master Class: Probity and Public Procurement | Madrid
■	30.6. – 1.7.	Master Class - “State Aid Uncovered” with Prof. Dr. Phedon Nicolaides | Florence

■	European State Aid Law	EStAL
■	European Structural and Investment Funds	EStIF
▨	European Procurement and Public Private Partnership Law	EPPPL

For more information and registration please visit our website

www.lexxion.eu

CoRe Blog

Official blog of the
European Competition and Regulatory Law Review

Let’s talk competition law!

Follow us now:
coreblog.lexxion.eu

The Legal Publisher
lexxion

AI Governance

Paul Nemitz, European Commission

Technology and (economic and political) power are entering into an ever closer symbiosis. A technology that knows more about man and the world than man knows about himself, and that is given ever more decision-making powers, leads to a massive asymmetry of knowledge and power in the relationship between man and machine. Classical models of action and decision-making in democratic societies are challenged by these developments.

The question of technical power and the control of technical power is raised in a new way. Who will decide in future? And, as Shoshana Zuboff asks, 'Who decides, who decides?'

When technology changes the power to shape things so radically, it is not surprising that the fundamental intellectual and cultural concepts on which modern societies are based are subjected to a stress test.

DELPHI – INTERDISCIPLINARY REVIEW OF EMERGING TECHNOLOGIES

Delphi is a pioneering interdisciplinary review of emerging technologies as seen through the perspectives of experts from the fields of science and technology, ethics, economics, business and law. Inspired by the idea to encourage inclusive, thoughtful – and sometimes unsettling – debates on the many opportunities and challenges created by technological progress, the international quarterly review brings together authors with different professional backgrounds as well as opposing views. Contributions to *Delphi* come in compact formats and accessible language to guarantee a lively dialogue involving both thinkers and doers.

ISSN 2626-3734