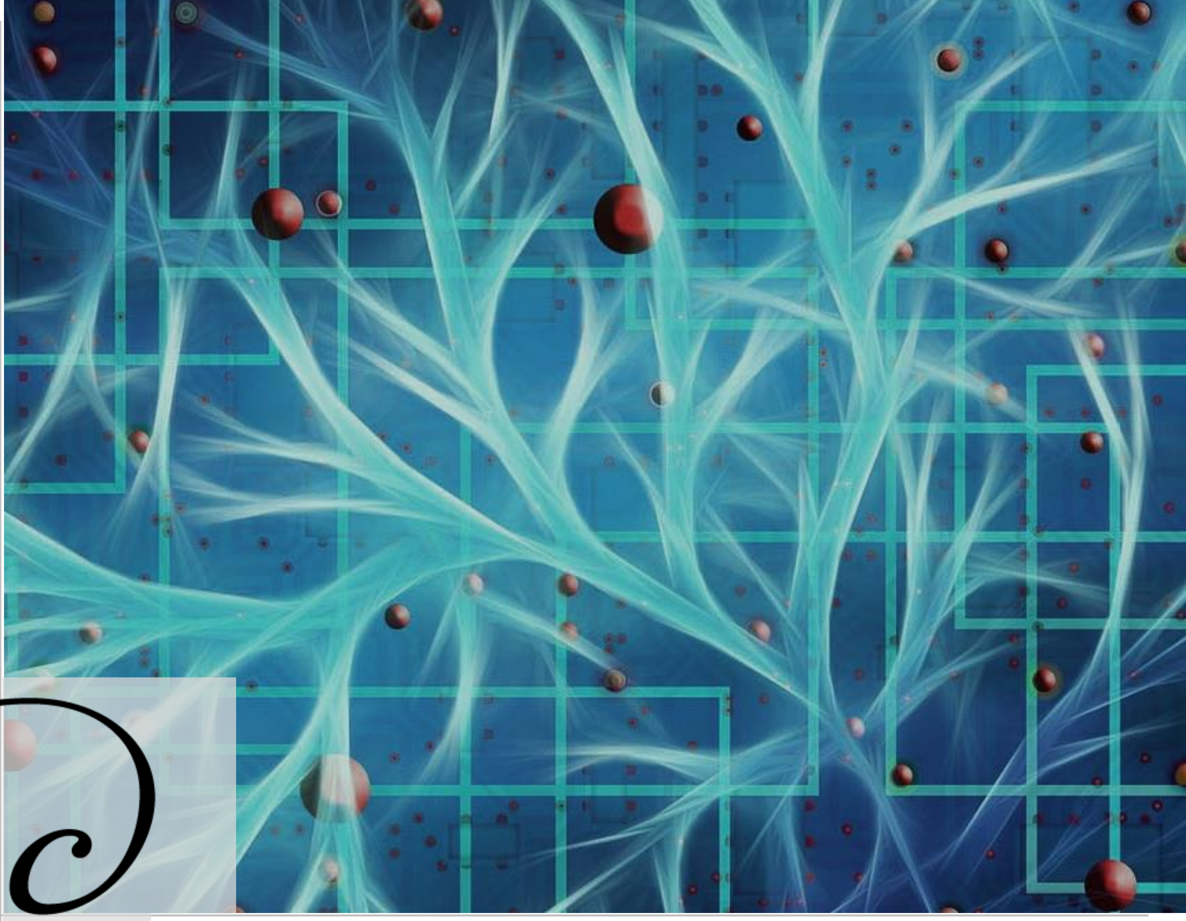


international risk
governance center

REPORT

The Governance of Decision-Making Algorithms

Inspired by a workshop held at the
Swiss Re Institute (Centre for Global Dialogue)
Rüschlikon (Zürich)
9-10 July, 2018

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Citation: EPFL IRGC (2018). The Governance of Decision-Making Algorithms. Lausanne: EPFL International Risk Governance Center.
DOI: 10.5075/epfl-irgc-261264

Available from: irgc.epfl.ch and irgc.org

For any questions, please contact: irgc@epfl.ch

Contents

LIST OF BOXES

PREFACE

Societies are becoming increasingly dependent on digital technologies, including decision-making algorithms applied across a broad spectrum of areas such as traffic, health, administration, insurance, commerce, news, advertising or weapons. As decision-making algorithms become more widespread and capable of (i) processing large bodies of information, and (ii) optimising choices for humans and institutions, they can bring crosscutting benefits to society. Yet, they are also increasingly complex and remain challenging to conventional decision-making where human judgment and ethical deliberation matter.

On 9-10 July 2018, with support from the Swiss Re Institute, IRGC organised a multi-disciplinary and multi-stakeholder workshop on ***the governance of decision-making algorithms.*** In a round-table setting, a group of 30 participants representative of AI and data science, regulation, policy analysis, industry and insurance were invited to discuss the challenges that arise around relying on decision-making algorithms and ways of governing them.

The specific questions guiding the workshop discussions are listed in the Appendix. Participants focused more specifically on challenges around the propagation of undue biases and challenges with designing fair decision-making learning algorithms; how reliable, explainable and accountable decision-making learning algorithms are; the attribution of liabilities when something goes wrong, general protection and good handling of data as well as cross-cutting concerns about trusting machines or systems that learn to make decisions instead of and on behalf of humans.

Inspired by the workshop discussions, this document summarises its highlights, followed by a more detailed elaboration of some specific issues raised. While the workshop considered the governance of decision-making algorithms more generally, this document brings particular attention to decision-making learning algorithms (DMLAs) which are for now more novel and rare, but probably of growing relevance for the future.

The document may be of interest to a varied audience of policy-making, business and/or research professionals whose domains are accommodating or contemplating the use of DMLAs. Its interdisciplinary character helps take stock of a broad spectrum of risks around DMLAs so as to advance their governance.

HIGHLIGHTS

Workshop participants raised and addressed various points around the governance of decision-making learning algorithms (herewith DMLAs). The following insights are worth highlighting.

1. **Technology and governance are tightly interconnected.** The governance of DMLAs entails both technical and non-technical aspects, and the challenge is to relate them well. An important part of governance by DMLAs will be to define desired policy, research and business goals in a way that allows engineers and developers to embed the governance rules or good norms into the very functioning of the algorithms. It is also necessary to include a mechanism of quality control (e.g. to check adherence to these rules or norms).

2. **The advent of 'learning' algorithms.** Amidst the different types of algorithms and different ways to define them, algorithms that *learn* and *self-evolve* warrant particular attention. While algorithms have already been making decisions, the key distinction is that these algorithms will no longer be "programmed" but increasingly "learned" and adaptive, giving them an ability to tackle tasks in domains that were previously done by humans trained or entrusted for such purpose. For the time being, however, there is limited take-up on DMLAs. Besides a handful of more dominant players (e.g. tech giants and certain governments) and industry innovators, most organisations are still exploring *what is possible*, to the extent that they are considering the use of DMLAs at all.

3. **There is a need to differentiate across applications and domains.** Amidst different possible modes and layers of governance, *addressing the governance of distinct and shared risks* – such as around undue bias and social discrimination, methodological inadequacies or shortcuts (e.g. inaccurate data inputs and outputs, learning what is undesirable, inappropriate repurposing of data, etc.), loss of accountability and of human oversight, as well as inappropriate or illegal surveillance and malignant manipulation – will be important to implement DMLAs and enable trust.

4. **Prior context and application domain matter for evaluating learning algorithms.** When DMLAs are deployed in specialised domains – such as in medicine, transportation, insurance or public administration – they do not develop in a contextual vacuum. It is worth recalling that there already are certain decision-making practices, analytical thresholds, prescriptive or historical norms in place that matter for evaluating or juxtaposing the performance of DMLAs vis-à-vis alternatives. An overarching question is how to evaluate learning algorithms against existing benchmarks in human decision-making, which are also not error or bias-free. When the benchmarks are lacking, how to define them? While the regulatory backdrop can vary by domain, specific applications of DMLAs require spelling out the relevant benchmarks against which we can evaluate their performance.

5. **The accuracy of DMLAs is not granted but must be more carefully thought-out.** Insofar as machine learning is increasingly a critical feature of algorithmic decision-making, the accuracy (or correctness) of DMLAs' outputs is what needs to be established, especially when the decision-making process is difficult to explain and/or the outcomes are difficult to interpret. Greater attention is needed to ensure more robust methodological practices (particularly regarding the quality and appropriate use of data) but also to probe the computational dynamics and learning at play. A particular concern is that we do not yet quite know how to test machine learned and adaptive algorithms. DMLAs may have the potential to reduce the number of errors in aggregate, but those errors may be qualitatively worse when judged against the expectations of equivalent

human decision-making. It is useful to design for potential failure (e.g. engineering 'better' worse-case algorithmic behaviours) but also increasingly necessary to decide, perhaps even regulate for, how we determine accuracy for DMLA systems.

6. **A key challenge with DMLAs revolves around biases.** While not all biases are problematic, many are. *Algorithmic bias* – at the level of data inputs, learning context or outputs – can yield discriminatory treatment along sensitive or legally protected attributes of race, gender, ethnicity, age, income, etc. Algorithmic bias remains tricky to address, particularly when it manifests through proxies. De-biasing techniques exist, but entail a trade-off: in order to evaluate whether undesired proxy measures creep into one's decision-making system, one would need more information and may have to include the very sensitive categories (of race, gender, age, ethnicity, etc.) to know if one has sufficiently minimised the bias. Amidst growing attention to new types and capabilities of algorithms, it is worth emphasising that obtaining more complete and unbiased data remains a pervasive and significant challenge.

7. **Under which circumstances can, or should, humans remain in control?** A key question when evaluating whether we make the right choice in relying on decision-making learning algorithms for specific applications is to ask if *humans are in the loop* (actively in control), *on the loop* (i.e. in alert mode, able to take control if need be) or *off the loop* (unable to take control back). While 'on' the loop may strike as the most balanced approach in that it suggests an ideal level of control, it comes with some risk that humans may struggle to 'jump in' when handed control if lacking the relevant context, practice, attention and time for making a critical decision.

8. **The development of standards, principles and good practices is needed** as they help industry, organisations and various developers embed best practices 'by design'**.** IEEE's Ethically Aligned Design Principles and standards (Global Initiative for Ethical Considerations in the Design of Autonomous Systems), the Asilomar principles for Responsible AI, or other initiatives by international organisations like the OECD, show that the development of governance arrangements for the programming, implementation and use of DMLAs is a shared concern.

9. **Defining accountability, responsibility and liability remain central.** It becomes ever more important to determine who assumes what responsibility in DMLA's development and use, what liabilities arise in case of erroneous or wrongful applications, and what are the concrete ways to demand and deliver accountability. Liability attribution is particularly challenging given the 'many hands' partaking in the design and deployment of DMLAs, but better defining legal uses of DMLAs can also help different organisations determine whether to enable or accelerate their adoption. Regulated sectors might be able to re-inscribe the deployment of DMLAs in their existing regulatory frames. New regulation such as EU's Global Data Protection Regulation (GDPR) enhances the maturity of the legal framework around permitted use and repurposing of data, data subject rights and other key privacy controls, but has not significantly developed the EU's regulatory framework around automated decision-making or the 'right to explanation'. There remains a general prohibition on making such decisions using personal data, but there are many possible exceptions to that prohibition and these are implemented in a patchwork manner across the European Union.

10. **Engineering digital trust remains a critical challenge and is increasingly relevant**. While it is possible to look for ways to mandate or improve trustworthiness, the challenge is increasingly about trusting the broader ecosystem in and around DMLAs for which a governance structure might help.

WORKSHOP SUMMARY

1. Setting the scene

Societies are becoming increasingly dependent on digital technologies, including decision-making learning algorithms (herewith DMLAs)[1] applied across a broad spectrum of areas such as:

>> transportation (e.g. autonomous driving),
>> health (e.g. diagnostics and prognostics),
>> research (e.g. analysing satellite data, medical research, etc.),
>> administration (e.g. predictive policing, criminal risk assessments),
>> surveillance (e.g. citizen scoring schemes, counter-terrorism),
>> insurance (e.g. processing claims, addressing insurance fraud, etc.),
>> news or
>> advertising.

Benefits

As DMLAs become more widespread and capable of i) processing large bodies of information at speed and at scale, and ii) optimising choices for humans and institutions, they bring many benefits to society. The benefits of DMLAs manifest themselves in a more cross-cutting fashion in terms of:

- **analytic prowess,** or analysing great volumes and flows of data, from multiple sources, in ways that are not possible for humans;
- **efficiency gains**, or generating outcomes more promptly and less costly than could be done by human processors;
- **scalability,** or the potential to bring in large domains to draw linkages, find patterns and/or yield outcomes, either in a specific domain or across multiple domains, geographic, substantive or otherwise;
- **consistent application** whereby DMLAs can process information more consistently and systematically than humans;
- **adaptability,** or the ability to process and learn with dynamic data and adapt to changing inputs or variables fast, sometimes in real time; and
- **convenience** or performing tedious or time-consuming tasks so as to free up human time for other meaningful pursuits.

In more illustrative terms,

>> In medicine, algorithmic decision-making can help with diagnostic and prognostics, especially around 'omics' (using genetic information for precision medicine) or with analysing medical reports;
>> In transport, algorithmic decision-making helps automate driving and could, with due care for reliability, improve driving convenience and safety;
>> In space research, it can help analyse publicly available satellite data and improve earth observation;
>> In administration, it can help automate repetitive and fatiguing tasks for humans (e.g. processing different claims, documents, etc.) as well as fight fraud, financial crime or identity theft.

[1] For working definitions of DMLAs, see section 2 (Which algorithms?)

Box 1: DMLAs in precision medicine

Precision medicine is an emerging field for disease diagnostic, prevention and treatment that considers individual variability in molecular data like genomics, environment and lifestyle. It aims to improve the health status of patients, and to help and support the prevention of the onset of a disease. The vast quantity and diversity of information (including heterogeneous large-scale biological datasets and clinical outcomes) that is required to make sense of a person's genetic, medical and lifestyle profile requires the use of artificial intelligence techniques to analyse the data. AI and machine learning are increasingly used for that purpose: to make diagnostics and prognostics (including for prevention), and to develop biomarker-based targeted therapies. Like for any application of DMLAs, issues for consideration include:

- How to keep private those data that should remain private and confidential? How to make sure that a person will not be discriminated, e.g. based on his/her genetic profile?
- How to work with all data, even confidential ones, that is necessary for improving the accuracy and reliability of the diagnostic, and informing decisions?
- How much automation should be authorised in the decisions that affect the health of an individual? To what extent are researchers, clinicians, medical doctors and patients comfortable that a DMLA is used to make decisions about a patient's health, considering that the complexity of the analytical process goes beyond what a human decision-maker can work with, and that there is and will continue to be uncertainty about the outcome of the analytical process?
- If a wrong medical diagnostic is made, using complex and large datasets of genetic and medical information, who is responsible?

Risks

At the same time, the development and use of decision-making learning algorithms can have negative consequences. Scholars and stakeholders identify important risks such as[2]:

- **Erroneous or inaccurate outcomes**, which can be difficult to identify or correct due to the intrinsic lack of transparency on the provenance of decisions, and difficulty to test DMLAs;
- **The problem of software correctness,** in that DMLAs are embedded in software and we lack sufficient knowledge on how to produce software that is always correct;
- **Threats to data protection and privacy**, as there is tension between privacy protecting rights such as 'the right to be forgotten' and the need for more complete and unbiased datasets for decision-making learning algorithms to live up to their potential;
- **Social discrimination and unfairness**, most notably through the reproduction of certain undue biases around race, gender, ethnicity, age, income, etc.;
- **Loss of accountability,** as some decision-making learning algorithms resemble 'black boxes' such that decision-making is difficult to understand and/or explain and thus difficult to account for any responsibilities or liabilities attending such decisions;
- **Loss of human oversight**, as decision-making learning algorithms are increasingly deployed in domains (e.g. of medicine, criminal justice, etc.) where human judgment and oversight matter;

[2] A different classification of risks identifies *manipulation*; *diminishing variety* (e.g. echo chambers, filter bubbles, biases and distortions of reality); *constraints on freedom of communication* (e.g. censorship through intelligent filtering); *surveillance* and *threats to data protection and privacy*, *social discrimination*; *violation of intellectual property rights*; *abuse of market power*; *effects on cognitive capacities and the human brain; growing heteronomy;* and *loss of human sovereignty and controllability of technology,* (Saurwein, Just, & Latzer, 2015).

- Excessive surveillance and excessive social control, as decision-making learning algorithms are deployed by powerful actors, be they governments or businesses, to survey citizens or unduly influence their behaviour;
- Manipulation or malignant use, such as for criminal purposes, interference with democratic politics, or in human rights breaches (e.g. as part of indiscriminate warfare).

In more illustrative terms,

> >> In medicine, algorithms can become highly proficient at learning different relationships or patterns, but they often cannot explain how or why these exist (Greenwald & Oertel, 2017). For example, a predictive machine learning model trained to predict the risk of death in pneumonia patients may also learn that patients with asthma have a better probability of survival, but it may struggle to explain this somewhat counterintuitive finding (e.g. that asthma was predictive of survival because the patients had been treated more aggressively) (Lipton, 2016).
>
> >> In justice, algorithmic decision-making can be deployed in ways that also discriminate against citizens on the basis of sensitive attributes of race, gender, health, age, etc. As per ProPublica's investigation, in criminal justice, software that predicts recidivism can perpetuate historically problematic racial biases[3]. Unfair and inaccurate outcomes can also result from learning algorithms that proceed with incomplete data or are not exposed to the full outcome for one large group (e.g. people denied bail, if they would commit a crime). Algorithmic predictive policing can also unduly target a 'type' of person/community (e.g. of particular ethnic belonging, race, immigrants, etc.) as more 'crime-prone', bypassing or ignoring important socio-historic dynamics or biases at play.
>
> >> In public administration, algorithmic decision-making can link biometric and other personal data to not only the provision of concrete public services (e.g. healthcare, education, etc.) but also to score and rank citizens in ways that control their behaviour. China's experiment in 'social ranking' raises serious questions about how governments mobilise the technology for improving the provision of public services while at the same time exercising social and political control (Hvistendahl, 2017).
>
> >> In national security, small algorithmic errors can prove big. For example, in the case of SKYNET (a system proposed to the NSA to help identify terroristic couriers based on smartphone data), even small false-alarm rates (0.008%) can be big, affecting some 4,400 individuals when the algorithm is applied to a very large pool of data (circa 55 million datasets (Zweig, Wenzelburger & Krafft, 2018).

Policy-makers face a difficult balancing act between allowing and incentivising the meaningful uses of DMLAs from the adverse ones. Some European leaders, for example, recognise the potential of DMLAs for key sectors – notably in healthcare – while noting also that 'this huge technological revolution is, in fact, a political revolution'[4].

[3] As per ProPublica's reporting on the COMPAS software used to predict recidivism with undue racial impact (Anwin, 2016) and reactions to it (Corbett-Davies, Pierson, Feller, & Goel, 2016)

[4] Interview with French President Macron (Thompson, 2018)

An additional challenge is that all aforementioned risks can also arise in traditional decision-making, with humans in control. Comparing the level of risk of DMLAs to the status quo is therefore further challenged by our understanding of current decision-making practices in often complex institution systems.

Given both potentially significant benefits and risks in using decision-making learning algorithms, the challenge remains to bring the two in perspective and capture their interplay by domain or specific application. Box 2 illustrates risks to consider along with potential benefits.

Box 2: Some examples of risks vs. benefits related to using decision-making learning algorithms

	Potential risk of relying on DMLAs	*Expected benefit of using DMLAs*
Insurance contracts	Incorrect actuarial analysis misprices risk or introduces unfair discrimination in prices	More efficient allocation of risk, e.g. through better actuarial analysis and fraud detection
Medical diagnostics/prognostics	Wrong medical diagnosis, prognostic or treatment decision	Improving the capacity to diagnose, prevent or treat life-threatening diseases
Automated driving	Wrong assessment of a car environment (car-to-car and car-to-infrastructure) leading to an accident	Benefits of autonomous (connected) guiding of vehicles, such as increased traffic efficiency and fewer accidents Comfort and convenience
Predictive policy		
- Criminal justice	Incorrect prediction of recidivism, potential unfair discrimination	Ability to enforce rules a priori by embedding them into code
- Public services / social benefits	Incorrect, potentially unfair discriminative distribution of social benefits	Embedding into code rules for a loan or social benefit attribution
- Face recognition (ID)	Undue or illegal citizen surveillance	Reducing eyewitness misidentification (a lead cause in wrongful convictions)

After noting the importance of analysing the distinct risks around decision-making algorithms, participants drew attention to a more fundamental question of what *types of algorithms matter the most*.

2. Which algorithms?

Following workshop discussions, decision-making learning algorithms can be understood as *information systems that use data and advanced computational techniques, including machine learning, to issue guidance or recommend a course of action for human actors* and/or *produce specific commands for automated systems.* While traditional algorithms are generally an outcome of a pre-specified engineering process, decision-making learning algorithms (DMLAs) differ in that their operation is an outcome of both input data and learning method.

Definitions of algorithms actually vary from the more technical or logical descriptions – as mathematical constructs[5], 'well-defined set of steps for accomplishing a certain goal' (Kroll and al., 2016), or 'sets of instructions to perform a task, producing an output from a given input' (Doneda & Almeida, 2016) – to more broad conceptions – as general purpose technologies (Bresnahan, 2010), even mediators and constructors of reality (Aguila-Obra and al., 2007). From a more critical standpoint, they can amount to

[5] A more elaborate discussion on ways to define algorithms in Mittelstadt, Allo, Taddeo, Wachter & Floridi, 2016

ideologies (Mager, 2012) or gatekeepers (Tufekci, 2015). For some, the definition of algorithms entails 'any large, complex decision-making software system **and** the larger environment in which it is embedded' (Smith, 2018).

This variance, or even ambiguity, in the definition of algorithms is not too surprising. We may expect further re-defining as both algorithmic capabilities/uses and our views of them keep evolving. As algorithms can select, classify, assign relevance, optimise, automate and/or undertake categorical decisions that yield concrete outcomes with consequences, the focus of the IRGC multidisciplinary and multi-stakeholder workshop was on algorithms with some *decision-making capacity*. Such algorithms go beyond diagnostics. They can yield prognostics on the basis of which organisations may subsequently automate a decision with important consequences for individuals. Examples include decisions about executing contractual arrangements (such as insurance claims), predictive policing and surveillance, various forms of individuals' profiling (such as for social benefits, loan distributions, advertising, etc.), or complex medical diagnosing and treatment purposes. **The more automated or 'independently' deciding algorithms are, the more they need to be scrutinized.**

Algorithms that *learn* from the data and optimisation processes to which they have been exposed and *self-evolve* (including through interactions with each other in networks, online, etc.) warrant particular attention. While algorithms have already been making decisions, a key consideration is that the algorithms are no longer "programmed" but increasingly "learned" and adaptive, which gives them the ability to tackle tasks in domains that were previously done by humans trained or entrusted for such purpose. Thus participants emphasized their growing salience, even if for the time being there is limited use of learning algorithms by businesses and governments, many of whom are still exploring what is possible.

Algorithms proceeding on the basis of pre-specified or fixed rules strike as more predictable, but 'once an algorithm is learning, we no longer know to any degree of certainty what its rules and parameters are' and 'we can't be certain of how it will interact with other algorithms, the physical world, or us' (Smith, 2018). As such, DMLAs can potentially behave in unpredictable ways –as Microsoft's chat bot Tay did (Angulo, 2018) – and also be intentionally misused. If 'learned behaviour' increasingly becomes a critical feature of algorithmic decision-making, learning methods and approaches differ enough to warrant more concrete questions such as when, where and how does learning happen[6].

While noting the growing importance of DMLAs, participants discussed why we witness a general need for governing the development and use of decision-making algorithms.

[6] With respect to machine learning algorithms, a basic typology distinguishes between *supervised* (outcome to be predicted from a given set of predictors, training continues with labeled data until model achieves certain accuracy), *unsupervised* (no predefined target or outcome to predict, no training with labeled data, but processing of information proceeds to find patterns, even ones not yet known), *semi-supervised* (in between the two) and *reinforcement* learning (continuous or iterative learning from past experience and the environment to arrive at or maximise 'best' outcome) (Fumo, 2017)

3. Why governance of decision-making algorithms?

In addressing the question of 'why' the governance of decision-making algorithms matters, the following insights transpired:

As technology can bring about both positive and negative changes, **societies have historically developed governance around distinct risks arising with emerging technologies.** The advent of railroads and trains, aviation, nuclear weapons and energy or nanotechnologies, to name but a few, brought about both governance measures and societal reflection on how to better evolve with them. DMLAs are not *a priori* exempted from this general need for governance. They too are subject to a line of basic questions, such as: Is the technology sufficiently accurate, reliable and fair? Does it improve on existing methods of analysing information and producing decisions? Who pays, who reaps the most benefit and who bears the most risk? What societal values are at stake? How vulnerable are they to ill or malignant use? What is their impact on existing socially consequential problems?

While DMLAs may learn fast and continuously and they can perform certain tasks more scalably, consistently or optimally than humans can, there are concerns that they struggle to proceed in ways that are reflexive and reflective of societal preferences, that they may make decisions that strike us as lacking common sense[7], that they cannot recognize important context switches, own errors, nor weigh the consequences of their decisions. **DMLAs remain particularly challenging to decision-making where the stakes are high and human judgment varies, but matters.** Perceived or potential lack of interpretive skills, common sense, self-awareness, empathy or social intelligence raises questions as to the type and extent of decision-making that can be delegated to algorithms, and with what human supervision (if any). There are particular concerns about delegating ethical decisions to DMLA-enabled systems as well as decisions revolving around more open, contested or ambiguous questions on human dignity, loss, or fairness. If such notions are socially contested and (re)negotiated, yet algorithms increasingly learn to undertake decisions that imply them, how to adequately define or embed what is fair, unbiased or socially 'good' in technical terms as well?

We may hypothesise that i) people may be more likely to accept a decision if they can empathise with the human decision-maker even in cases where that decision-maker has previously made *subtle* or obvious errors, and ii) less likely to accept a decision if it is made by a DMLA that has previously been shown to make *obvious* errors in its decision-making, even if overall the decision-making record of the DMLA is measured as better than the equivalent human. **A key challenge is the governance and acceptance of algorithms that are also imperfect**. Especially challenging are rejections of techniques and advances at the first failure – a risk which may need distinct governance measures such as public benchmarks and certifications.

Insofar as algorithms become increasingly part-and-parcel of governance itself (e.g. in predictive policy-making, boardroom governance, etc.), participants recalled that algorithms cannot hold all the governance nor are they the only vehicle for it. Governance in our times is rarely unitary or one-dimensional (see Box 3 for a broader discussion). **Given different governance modes, levels and approaches, algorithms cannot miraculously absolve or displace such multiplicity, but must also contend with established or additional layers of governance**.

[7] The Todai Robot was designed to pass a university entrance exam but struggled with common sense, as Noriko Arai, the lead scientist on the project, explains in her Ted talk (Arai, 2017).

The peculiarity of DMLAs as a technology is that their governance entails both technical and non-technical aspects. The two are deeply interrelated, and the challenge is precisely to relate them well so that salient social preferences are reflected in computational processes and, in turn, computational prowess is used to improve on socially consequential problems. In section 4, we review technical tools and considerations that workshop participants flagged in an effort to better engineer DMLAs governance.

A recurring workshop insight was that although machine learning is a general-purpose technology, there is important variation as to where, when, why and how DMLAs are deployed, which warrants a more differentiated approach for the governance of distinct risks that accrue.

Box 3: General governance considerations

While it is possible to elaborate on governance in more scholarly terms (Bevir, 2009), for a more varied audience, it is useful to begin by noting that governance is not equivalent to government, although governments very often partake in it (most notably by way of laws, attending oversight, regulatory bodies, etc.). There are different actors – other than governments – that partake in framing and practising governance and governance is applied to different issue-areas as well.[8] Governance entails an array of possibilities (Saurwein and al., 2015), including state regulation, international laws, private regulation (self-organisation at the level of individual organisations), collective self-regulation (e.g. at the level of sector or industry, by way of standards, ethics boards, etc.), co-regulation, civil initiatives but also certain anticipatory approaches such as through responsible research and innovation (Stahl, 2013).

In their distinct ways, scholars of politics and law have for example drawn distinctions between 'hard' and 'soft' law (Shaffer & Pollack, 2009)[9], or 'hard' and 'soft' power (Nye, 2009). It is not altogether evident how 'hard' or 'soft' governance measures may be separated in the case of DMLAs. As one participant noted, '*within the term governance, we can think about a broad configuration of legal, ethical, and professional behaviours or conventions that, taken together, guide the development and use of data and decision-making algorithms*'.[10]

While there are already prominent tropes in the conversation around governance of DMLAs (such as intensified calls for algorithmic transparency, algorithmic impact assessments, fairness, explainability, etc.), questions abound about what matters the most. Questions such as whether to govern by way of a 'global default' (Wagner, 2016), establish new independent bodies or rely on existing ones[11]; regulate at the level of code or operators ("How Policymakers Can Foster Algorithmic Accountability", 2018) (Stahl, 2013); pursue explicit international standards or more implicit research and development norms (such as Responsible Research and Innovation in the EU); better define explicit user rights (e.g. on privacy, right to explanation, etc.) or emphasize digital literacy more; embed democratic principles in DMLAs or take care that the technology itself does not become a new international fault line[12]; etc. In other words, increasingly important for the governance of risks around DMLAs is that the relevant trade-offs are more explicitly exposed and that governance arrangements emerge through processes recognised as legitimate.

[8] We speak about multilateral governance, corporate governance, internet governance, risk governance, etc.

[9] There is also a critique of soft law (Klabbers, 1996)

[10] Participant comment

[11] See The IEEE Global Initiative on Ethics of Autonomous and Intelligent System's discussion on 'lack of an independent review organization' in Ethically Aligned Design: *A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems,* V2 (2017: pp.70)

[12] See "The battle for digital supremacy – America v China", 2018

4. Governance of distinct risks

By both technical and non-technical means, the governance of risks around algorithmic decision-making must take stock of important variations as regards:

1. **The underlying method** by which algorithms arrive at decisions, in particular where does learning take place, with what datasets, with what human guidance and in what environment;
2. **Human input in and control over algorithmic decision-making**, namely when are, metaphorically speaking, **humans in, on** or **off the loop** , and what it means to monitor undue errors or failures, and if necessary, to contest algorithmic decision-making;
3. **Domain-specific deployment,** in particular what pre-existing governance practices or regulatory context in the sector in which DMLAs are deployed and what key benchmarks or salient norms against which to assess the quality of algorithmic decision-making; and
4. **The purpose and people behind algorithmic decision-making,** namely the risks around malignant or nefarious use, but also that the technology benefits specific organisations or actors at the disadvantage of larger segments of society.

Elaborating further:

4.1 Risks around the underlying methodology: or how accurate, reliable and explainable are DMLAs?

With respect to how algorithms can produce decisions, key questions are: with what kind of data/datasets DMLAs work and how to handle privacy and biases while pursuing accuracy; what is the learning context and what kind of learning takes place and where; and to what extent can and should algorithms explain or account for their decisions.

While much attention goes to algorithms' growing computational capabilities (e.g. advances in deep learning, etc.), that is only half of the story: **sophisticated code without the right data may not live up to its potential**[13]. When data is increasingly viewed as the new 'oil'[14], concerns around its availability, what personal data is/isn't accessible to algorithms, users' consent and re-purposing by various other third parties naturally grow. But the challenge of more complete and unbiased data must not be understated through mere insistence on algorithms' growing sophistication.

Accuracy

While privacy remains paramount and pursuable through privacy-preserving techniques, it is also important not to lose sight of risks around inaccuracies that may yield errors, bad decisions or misrepresentations. To this effect, **workshop participants noted the importance of improving the quality and representativeness of data**, such as by adequately curating data for use (vs. relying on 'raw' data), with due attention to not only *privacy* restrictions, but also *undue biases*, susceptibility to *errors*[15],

[13] The case of IBM's Watson Health revealed that lack of adequate training data can undermine the potential of learning systems in healthcare. "If Watson has not, as of yet, accomplished a great deal along those lines, one big reason is that it needs certain types of data to be 'trained'. And in many cases such data is in very short supply or difficult to access. That's not a problem unique to Watson. It's a catch-22 facing the entire field of machine learning for health care." (Greco, 2017)

[14] There are popular depictions of data as the new oil ("The world's most valuable resource is no longer oil, but data", 2017), even if data might not be scarce or costly to extract as oil

[15] E.g. spurious linkages or small noise confuses the data, leading to different outcomes.

over/under-fitting issues[16], and exposure to *randomness*[17]. As an illustration, automated driving will not improve on safety and reliability if it lacks sufficient training data, if it does not pick up on common human mistakes[18], or if the learning is primarily from one type of setting[19] or an insufficiently varied population of drivers. With respect to predictions of recidivism and bail decisions, inaccurate results may be obtained if the training algorithms are not exposed to the full outcome for one large group (e.g. likelihood to commit a crime for people denied bail)[20].

Therefore, recognizing that data quality and representativeness matter for the validity of algorithmic decision-making, participants further noted **the trade-off between restricting the availability of data and achieving better accuracy**. For example, removing or anonymising certain data (e.g. to assure users' privacy, consent, etc.) may potentially lead to a less accurate dataset on which algorithmic learning and decision-making proceeds. This trade-off is particularly manifest in de-biasing techniques, as discussed below.

Accuracy, at the same time, also requires some interpretive nuance. The 'data' with which algorithms work are often around different people's behaviour, and yet what is accurate for one group of the population might not be for another. This is already more evident in the case of facial recognition technologies (Reilly, 2018), but it matters also for medical diagnosis and treatments, attribution of criminality to individuals and/or communities, or automated driving. Accuracy may have to be broken down, and a simple score may be deceptively complex to interpret. For example, an algorithm using past healthcare claims to predict users' number of hospital visits in a year may have high accuracy overall, but without proper contextualization in interpreting such results, healthcare providers or insurers may unduly target the elderly or subpopulations with certain pre-existing conditions (Tramèr and al., 2015). O'Neil's recounting of how complex algorithmic computations to measure the educational progress of students in specific districts, calculate how much of their advance/decline can be attributed to their teachers and ultimately distil a socially complex consideration into a simple score that informs which teachers are to be dismissed is revealing of how seemingly sophisticated computations can nevertheless yield suboptimal, even unfair outcomes such as dismissing capable teachers from schools and/or incentivizing teachers to game or prioritize metrics over students (O'Neil 2016). The overall analytic robustness of DMLAs depends on how accurately they handle data throughout the optimization or learning process – from drawing on data as 'inputs' to generating further data and decision 'outputs'. Yet in the process of learning and optimizing, DMLAs may also exclude information deemed irrelevant and may inadvertently reduce the diversity of information that users and operators actually encounter. Reduction in diversity of opinions may also result in 'poor' collective decisions or behaviour[21].

An overarching concern is how to evaluate DMLAs' accuracy against the human baseline. DMLAs may have the potential to reduce the number of errors in aggregate, but those errors may be qualitatively worse when judged against the expectations of equivalent human decision-making. Another challenge is that we do not always know how to test machine learned and adaptive algorithms. It thus becomes necessary to decide, perhaps even regulate, how we determine accuracy for DMLA systems.

[16] E.g. algorithms tailored too well to an existing set of data but struggling when confronted with information to which they were not previously exposed

[17] E.g. to not struggle in unexpected environments for which algorithms may have not been trained.

[18] E.g. not respecting stop signs, not using turn signals properly, etc.

[19] E.g. a particular geography, how urban or frequented the setting, etc.

[20] On the methodological intricacies of comparing bail decisions of judges to machine-learned algorithms of what a defendant would do if released, see Kleinberg et al., 2017.

[21] E.g. if all oil-trading companies train on the same data sets or points, at a given moment in time, there is some chance that all market players will sell/buy same assets, thus disrupting healthier market exchanges.

Ultimately, one must recognise human fallibility on the one hand and the impossibility of error-free algorithmic decision-making on the other. DMLAs can learn and process information in ways that differ from humans, and err differently too. There is also the risk of rejection of techniques and advances at the first failure. While neither type of decision-making – human or algorithmic – is error- or failure-proof, it may be possible to engineer safe-fails in a decision-making system. **Thus, designing for potential failure, or engineering for 'better' worse-case algorithmic behaviours might prove as important as improving overall behaviour.** Setting certain 'guard-rails' or boundaries which algorithms cannot cross and demand the ability to prove those boundaries as a matter of technical design can help render DMLAs more reliable over time.

Biases

Critical challenges related to quality data and algorithmic decision-making revolve around the so-called **problem of *algorithmic bias***. Biases can manifest themselves at different points or aspects, most notably in the data itself (directly or indirectly by way of proxies), the manner in which they are processed/optimised, and also at the level of interpretation.

Biases in algorithmic decision-making are often juxtaposed to human biases. Various evidence suggests that human decision-makers are often unduly biased and humans can also have good or bad days. It seems, in this respect, that biases are present in both types of decision-making. It remains debatable if humans can become better interpreters and correctors of their private and institutional biases, or if it is possible and desirable to pursue computational solutions that proactively detect and unlearn problematic biases over time.

While not all bias is 'bad' per se, **biased representations in data, learning context/method and outputs can lead to subtle or blunt forms of social discrimination**. The US Department of Housing and Development has, for example, filed a complaint that advertising algorithms on Facebook violate the Fair Housing Act as they discriminate against protected groups on the basis of disability, gender, race, ethnicity, age, etc. (Booker, 2018). What is further troubling about algorithmic bias is that when it facilitates undue discrimination, it may be systematic but not easily detected. For example, if biased machine-learning algorithms process the major bulk of resumes or decide to whom to advertise what jobs, they may unintentionally discriminate against applicants but the users themselves may not be aware that they are being discriminated against (Carpenter, 2015).

Particularly challenging are proxies, i.e. elements or information that embed certain biases implicitly[22]. Information like postal codes can proxy for more sensitive attributes such as race, gender, income, etc. When Staples, for example, implemented a differential pricing scheme for online purchasers whereby those closer to a competitor (e.g. Office Depot) received discounted prices, its promotion had a negative disparate impact on low-income customers residing further away (Valentino-DeVries, Singer-Vine, & Soltani). The challenge with proxies is that on the one hand, excluding a protected category does not mean that a proxy for it is not being created and on the other, eliminating proxies is problematic if they also contain other useful information.

De-biasing techniques exist, but there is an almost paradoxical situation at play: in order to evaluate whether these proxy measures creep into one's decision-making system, one would need more information and may have to include the very sensitive categories (such as protected categories of

[22] Generally speaking, a proxy is someone or something representing another. In the case of algorithmic decision-making, it has been defined as a 'feature correlated with a protected class whose use in a decision procedure can result in indirect discrimination' (Datta, Fredrikson, Ko, Mardziel, & Sen, 2017)

personal data) to know if one has sufficiently minimised the bias. Is this socially acceptable and does it require a legal or doctrinal shift?

Learning process

Besides attention to the underlying data, various considerations about the learning process are important.

Where does the learning happen? Knowing if the learning happens in the lab or in a real-world setting, and if the sector is regulated or unregulated can also inform the governance of DMLAs. Algorithms can be released in versions *when and after* important new input data has improved the algorithm, which has been tested before the release. In this case, the learning risk is partially governed through licensing-like processes. In contrast, when the learning takes place more frequently so as to facilitate dynamic or adaptive decisions, this creates DMLAs that are more difficult to test and audit, with a view to identifying unwanted learning or outcome. This is particularly the case when the developer and the user are the same entity.

Also worth noting are increased interconnections and interactions between different algorithms or applications where DMLAs are used. Algorithms could be confusing or contravening each other too, or evolving as so-called 'spaghetti code', which can prove problematic like in the case of a Toyota Camry driver killed by unintended acceleration[23]. More generally, as learning algorithms designed in one place for a specific purpose (e.g. recognition of objects or faces) interact with other learning algorithms for other purposes (e.g. in medical prognostics, security, etc.) to yield concrete decisions/outcomes, **the learning environment also grows more convoluted**.

Privacy

New regulation like the EU General Data Protection Regulation (GDPR) attempts to regulate the legality and appropriate use of data. Such regulation is often 'principles-based', backed by more prescriptive rule-sets to embed those privacy-protection principles into law. Even though principles-based regulation is often more adaptable to new technology, the workshop did identify some particular challenges with DMLAs.

For example, the GDPR contains principles to ensure that the processing of personal data adequately matches its stated purposes ('purpose limitation') as disclosed to the data subject ('transparency'). These purposes must typically be recognised and stated to data subjects up front. However, especially where DMLAs are used, the purposes for which the decisions are made may change as a function of the learning of the algorithm, for example where DMLAs are able to highlight hitherto opaque patterns. This may be by design of the technology, or more likely by an act of the operator, who may extend the use of the DMLA beyond their originally stated purposes when it becomes apparent that the DMLA may be somewhat suitable for the new purpose. **Without clear governance and controls, there is a risk of unchecked extension of the purposes of DMLA – where processing personal data – could cause a breach of GDPR**.

It is important to discuss data minimisation in any consideration of privacy risk. The principle of data minimisation requires that personal data be adequate, relevant and limited to what is necessary for the relevant purposes of processing. DMLAs often perform better with larger datasets from which algorithmic insights can be inferred. Does that mean that all data there is relevant and necessary? Or should those terms be taken to imply a threshold of relevancy and necessity, below which data is not

[23] Bookout and Schwarz v. Toyota case, following a September 2007 unintended acceleration event with a 2005 Toyota Camry that killed Schwarz and injured Bookout. (Smith, 2018)

retained? If the latter, how does one apply such a threshold given that the relevancy and necessity of particular data may be identified by DMLAs only in hindsight – after they have observed enough data to be able to learn from it. This may present a risk both to privacy and to the adoption of DMLAs as organisations grapple with how to satisfy their own regulatory compliance obligations.

Accountability and explainability

The extent to which DMLAs can account for their decisions/outcomes matters for understanding the provenance of important decisions and enabling some form of general accountability, even redress in case of specific wrongful/erroneous decisions. Accountability is a broad concept that can be linked to other notions such as transparency, due process, fairness, or legal responsibility. In the case of DMLAs, it is also more particularly interlinked with explainability – the degree to which it is possible to give an account of how decisions are made.

Explainability is a desirable but challenging feature, and potentially manipulated too. It is not necessarily presumed that human decision-makers explain themselves in any regular, complete, consistent or accurate way for all the decisions they undertake or in which they partake, nor that users and publics will always require and listen to elaborate explanations of decision-makers. Yet when something is new, uncertain or in the process of changing established practices, having some form of explainability might help establish certain baseline understanding about how certain critical decisions come about and reduce the risks of wrongful or erroneous decisions accruing with time. **Participants, however, also noted the risk of 'gaming it' by way of some parallel-construction** that mines a 'conveniently plausible' explanation (e.g. 'adversarial explainability', p-hacking) **as well as the challenge that feedback often benefits an adversary** (i.e. any feedback from an algorithm might help those who are attacking it).

Trade-offs can arise between pursuing explainability – such as by devising audits or reporting mechanisms, or restricting machine learning to work with more easily interpretable methods – and pursuing computational prowess through more fluid or fast-evolving decision-making systems[24]. **When explainability and performance have an inverse relationship, then assessing the impact on end-users and distinguishing high-stakes decisions from low-stakes ones may help define the degree and mode of explainability to enable**. For example, if algorithmic decision-making leads to a sustained positive impact on the life or livelihood of people, then with good performance, explainability becomes less of a requirement over time. For negative but significant outcomes, however, sacrificing performance for higher explainability is more salient. The underlying norm is that people should be able to know where a decision affecting them is coming from. The explanation may at times be more 'local' (e.g. how a decision for a particular entity was made) or 'global' (e.g. how various factors are weighed and interlinked to produce a type/class of decisions). It may also be counterfactual, for example attempting to isolate the factor that made a difference between obtaining a loan or not[25]. From an end users' perspective, the explanation must be intelligible (e.g. it does not suffice to simply put some code in the open).

Distinct, but closely related to explainability of algorithmic decision-making is the question of transparency about the choice of underlying methodology as well as the organisations' manner of dealing with demands about outcome, in a way that is open and accessible (such as by making their decisions verifiable or accessible for independent assessments). Insofar as the digital space is perceived as

[24] In general, 'AI systems do not automatically store information about their decisions. Often, this feature is considered an advantage: unlike human decision-makers, AI system scan delete information to optimize their data storage and protect privacy' (Doshi-Velez and al., 2017)

[25] Or 'local counterfactual faithfulness' (Doshi-Velez and al., 2017)

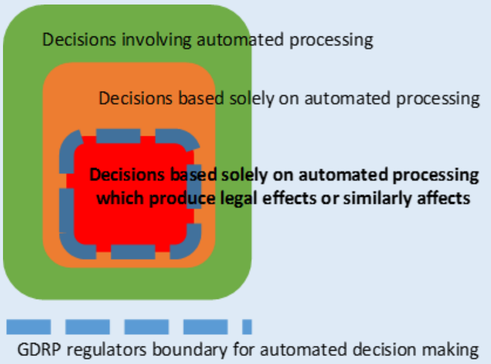
dominated by a handful of players with privileged access to data and better computational know-how (such as GAFA[26] and certain governments), a lack of transparency as to how such organisations curate the data they use for specific goals can also render any explanations of outcomes suspect**. For entities that for competitive reasons privilege commercial secrecy, explainability may be seen as undercutting it**.

Along with some explainability of outcomes, additional layers of transparency (about the data input, learning process but also when DMLAs are legally deployed) remain important for rendering algorithmic decision-making more accountable. While it resonates at a general level, the quest for transparency does not seem to be that straightforward as it plays out in many dimensions. There may exist socially accepted reasons and legal grounds for why users or organisations care to protect private data, proprietary code, etc. Disclosures about access to private data and introducing more independent monitoring platforms to verify or certify the quality of the underlying data (e.g. assess legality, accuracy, etc.) can help maintain a more trusted relationship with the data subjects. At the same time, further clarity on when there is a 'right to explanation' and a right not to be subject to automated decision-making (see Box 4) can also partially address the public's need to better understand the uses and limitations of DMLAs.

Box 4: The right to explanation and exemptions from automated decision-making in GDPR

In theory, the EU GDPR (2016/679) also makes some provision for new kinds of rights, notably the right not to be subject to automated decision-making and the right to an explanation.

As per Art. 22(1) *"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."* The red zone in the figure below shows that the scope of regulation may be more narrow than imagined.


Decisions involving automated processing
Decisions based solely on automated processing
Decisions based solely on automated processing which produce legal effects or similarly affects
GDRP regulators boundary for automated decision making

Source: participant presentation

As per Art 22(2), the general prohibition does not apply 'if the decision a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; b) is authorized by Union or Member State law...; c) is based on the data subject's explicit content.' This leaves some ambiguity as to when the right not to be subject to an automated decision based on personal data can/will be dis-applied as it depends, among others, on how different member states interpret it.

In a similar vein, a right to explanation appears in Article 13 whereby data subjects must be provided with information concerning "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."

[26] Google, Amazon, Facebook, Apple

However, there is no scholarly consensus on whether 'the right to explanation' can actually be realized (Wachter, Mittelstadt, & Floridi, 2017) (Doshi-Velez and al., 2017) (Malgieri & Comandé, 2017).

The law, as one participant noted, 'is far from settled'. Its interpretation and compliance with it may vary by countries and/or domains. Or as one legal scholar put it, 'the GDPR can be a toothless or a powerful mechanism to protect data subjects dependent upon its eventual legal interpretation… Supervisory authorities and their future judgments will determine the effectiveness of the new framework' (Mittelstadt et al., 2016).

With historical hindsight, various technologies have gone through a process of becoming more reliable, but also intelligible and transparent to the publics deliberating on them. The introduction of automated elevators at the turn of the 20th century provides some insights on how to improve the introduction of new, automated technology. When elevators stopped being operated by drivers (who manually opened/closed the doors and accelerated/ decelerated the elevator), the public was not particularly receptive of the change. The engineering of concrete measures –namely, an emergency stop button (allowing a human in the loop), an emergency telephone (to demand assistance and redress) and a calming voice aid curating the ride and explaining the working of the system – helped reduce risks of operator failure and assuage public fears.

DMLAs are much more complex than elevators, but thinking about tangible ways of enabling explainability and redress in case of failure, designing guardrails that algorithms cannot theoretically trespass as well as enabling human oversight in critical contexts could make them more reliable and socially palatable.

Besides probing and improving the underlying decision-making methodology, the workshop participants also discussed the role of human input in algorithmic decision-making.

4.2 Human control and algorithmic decision-making: when are humans in/on/off the loop?

Amidst general concerns about whether DMLAs augment or replace humans, it is important to recognize the different societal stakes and impacts, including the possibility that existing highly skilled professionals may reject these systems if they perceive that DMLAs threaten their status and/or livelihood. Aside from addressing the question of human control from a sociological and economic standpoint, in asking what is the appropriate human input in algorithmic decision-making, it becomes both conceptually and pragmatically relevant to set certain expectations as to when we expect DMLAs to mimic human judgment, when to aid it, correct it, or perhaps override it, and when we expect humans to be able to override them instead.

One way to reframe the question of human control vis-a-vis algorithmic decision-making is by asking when humans are:

- In the loop, e.g. humans are fully in control, in the sense that, at some stage in the decision-making process, the algorithm stops, hands the decision over to a human, who then instructs the algorithm to proceed in a certain way;
- On the loop, e.g. humans can take control if need be, in the sense that the algorithm informs a human supervisor who can intervene if need be to modify the decision or to take control (get back in the loop); and

- **Off the loop,** e.g. humans cannot take control, either because there is no supervision by design (in hypothetical fully automated self-driving cars), or because decisions are irreversible (launching nuclear missiles).

In practice, many “on the loop” systems put the human “off the loop” through the sheer amount of decisions and data being produced by the algorithm, making human supervision all but impossible (for example in high-speed trading where an algorithm can order complex trades worth millions in a fraction of a second). Thus while ‘on’ the loop may seem like the most balanced if not ideal ‘default’ option, it comes with some risk that humans may struggle to ‘jump in’ when handed control if lacking the relevant context, practice, attention and time for making a critical decision.

The discussion around being in/on/off the loop is more metaphorical than a precise typology (see Box 5 for levels of automation in driving), but it is a useful general heuristic to gauge i) when, how and with what latitude we can reasonably allow algorithms to take on decision-making attributes and ii) correspondingly, who assumes responsibility (moral, but especially legal) in case of decisions that lead to some harm.

Box 5: Driving automation

Automated and connected vehicles are equipped with dozens of sensors, radars, cameras or LIDAR's, which provide data to a central processing unit, which is itself connected to ‘infrastructure’. Thousands of signals are thus processed in real time to analyse road conditions and other factors. This requires data analysis and fusion, performed in the car processing unit and based on machine learning. The output is in the form of: *diagnostic* (e.g. monitoring the vehicle environment, information to the driver about road and conditions), *prediction* (e.g. instruction to the driver that he/she should steer the wheel or brake in order to avoid an obstacle), and *decision* (e.g. control of the steering wheel and brakes).
The level of automation and decision that is possible and authorised has been codified in five levels of driving automation systems (cf. On-Road Automated Driving (ORAD) committee, SAE 2016):

- **Human in control:**
  Level 0: no automation
  Level 1: driver assistance
  Level 2: partial automation. Control of functions such as steering or accelerating/decelerating may be delegated to the software, but the driver is and must remain in complete control of the vehicle at all times
- **Human can and must take control if need be:**
  Level 3: conditional automation. The vehicle can make an informed decision for itself, such as overtaking slower moving vehicles. However, human override is required if the vehicle is unable to understand a particular situation or to execute a necessary task, or if the system fails. Therefore, the driver must always, at any time, be alert and able to regain control.
  Level 4: high automation. The vehicle should be able to intervene if things go wrong or in case of a system failure. The human driver is not needed in most situations.
- **Human out of control, and cannot take control back**:
  Level 5: full automation. Human driving is eliminated. Human ‘driver’ is not in control. No task may ever require human attention. Human intervention may even be impossible.

Recent accidents caused by automated cars equipped with driver's assistance or auto-pilot have demonstrated the difficulty of the machine learning process used to facilitate automated and autonomous driving. In the case of an Uber accident in March 2018 killing a 49 year old-woman walking a bicycle at night, “*the self-driving system software classified the pedestrian as an unknown object, [then] as a vehicle, and then as a bicycle with varying expectations of future travel path[...] Only 1.3 seconds* *before impact, it determined that an emergency braking manoeuvre was needed. According to Uber, emergency braking manoeuvres are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behaviour. The vehicle operator is relied on to intervene and take action. The system is not designed to alert the operator".* (NTSB Preliminary Report 2018)

The case illustrates the complexity of determining if and when human drivers must remain in the loop (in control) and the extent to which they can be on the loop (back in control if necessary). Levels 2 and 3 of the SAE classification are the most difficult to translate in reality.

This issue around what counts as algorithmic decision-making and when are humans on, in or off the loop comes with important implications, notably around liabilities (see Box 6).

Box 6: Liabilities and algorithmic decision-making: the case of autonomous vehicles

What remains challenging in defining liabilities around algorithmic decision-making is the multiplicity of parties involved (users, programmers, operators, data brokers, new decision-making entities like algorithms themselves, etc.) and thus the difficulty of attributing intent and/or responsibilities in ways that are proportionate and effective among them. The case of a Toyota Camry unintended acceleration revealed how challenging and resource-consuming it becomes to establish what in the first place went wrong ("Toyota Unintended Acceleration and the Big Bowl of "Spaghetti" Code", 2013). The more complex and convoluted DMLA's computational processes, the more challenging to apportion responsibility to a particular segment of code, party, etc. Moreover, when explanations for an erroneous outcome are not technically forthcoming or straightforward, it encourages deniability almost by default (i.e. easier for different parties to pursue deniability until/unless proven otherwise). Regulators too might be tempted to pursue a regulatory shortcut and/or concentrate legal responsibility (e.g. insist that liability primarily accrues to the insurers with most expertise at spreading risks across society).

In the case of autonomous vehicles, it is not always easy to draw the line as to when responsibility for decisions rests distinctly with the developers (depending on whether software is deemed as service or product), manufacturers (if software has been embedded in products, hence product liability applies), drivers, owners, sellers or ultimately, insurers. For the time being in the UK, the Government has provided that liability should take similar form to non-automated car insurance for drivers, i.e. the insurer pays for damage for an accident caused by an automated vehicle[27], but may claim 'contributory negligence' from other parties (e.g. driver/owner, seller, manufacturer, transport system operators, etc.). What is novel is that where demonstrable and for an accident resulting from unauthorised software alteration by the user or from failure to update safety-critical software, the insurers' liability may be limited or excluded. Being able to prove the integrity of DMAs or DMLAs involved in automated vehicles may, therefore, become a key requirement of law enforcement, insurers and other affected parties, for example requiring evidential chain of custody and forensic examination of the datasets and software when investigating accidents.

While the multiplicity of actors makes it challenging to unchain or isolate key responsibilities and liability as legal concept and practice might itself adapt, a way forward is to evolve legal responsibility in more domain-specific terms. For example, to what extent is the sector in which DMLAs are being deployed already regulated and what are the relevant benchmarks or standards, which, if missed or unmet, open up a space for attributing legal responsibility?

[27] Provided driving took place in Great Britain, the vehicle insured at time of accident, damage caused to some person

4.3 Contextualization by domain and/or application

A recurring workshop insight is that DMLAs do not arise nor operate in a vacuum or clean slate per se: the application domain matters. They are increasingly deployed in specialised domains – like medicine, transportation, insurance, public administration, to name a few – where there already are certain governance practices, prescriptive norms, important benchmarks as well as some history and recognition of grave errors or failures. The question that arises is what the critical criteria or benchmarks are, in terms of expected risks and benefits against which a specific DMLA deployment can be calibrated.

> >> In medicine: Is, for example, the relevant benchmark or 'gold standard' in clinical research and diagnostics the 'truly positive case'? If so, how is this possible for DMLA enabled diagnostics? Should the latter's performance equal or surpass the current standard in order to be authorised?
> >> In transport: Is the 'fail-safe' option the standard that matters most and if so, can we mandate it for automated vehicles?
> >> In aviation: Should DMLAs be as reliable and trustworthy as current avionics systems?
> >> In criminal justice: is 'due process' critical for practising modern criminal justice and if so, can algorithmic decision-making embed this when attributing criminality or determining bail conditions?
> >> In predictive policy-making: Is 'neutrality' with respect to one's gender, race, ethnicity, etc. the relevant norm for non-discrimination in society and if so, can and should we de-bias DMLAs by using these explicit attributes?

In sum, **the regulatory backdrop and benchmarks against which we calibrate DMLAs can vary by domain or use, but they matter.** Their absence, as in emerging or unregulated sectors, does too**.** Medicine, justice, transport, finance, etc. are by now specialised and fairly regulated practices but also traditions, with their set of recognised benchmarks, best practices as well as memories of grave failures and checks against which algorithmic decision-making can be calibrated. Unregulated practices or domains, on the other hand, may involve more open multi-stakeholder conversations around articulating not only the applicable benchmarks, but also more adaptive yet risk-aware governance approaches for a technology and ecosystem in the making.

It also remains important not to lose sight of the bigger picture, namely to recognise the purpose and the people behind the deployment of DMLAs in different domains and probe in a more critical way to whom the benefits accrue most, who bears what risks, what are the incentives and what are the risks of adversarial use. As one participant reminded us with the words of Melvin Kranzberg, '*technology is neither good nor bad, nor is it neutral*'.

4.4 The purpose and people behind DMLAs: Who benefits? Whom or what to trust?

While those who may benefit from DMLAs are not only 'powerful' actors, there remain poignant, even politicised concerns about the motives of entities that develop or use them most.

Informational asymmetries – between data subjects, brokers, companies or various platforms where data is gathered – may also affect popular perceptions as to whether DMLAs are being put to good general use. Concerns about the direction of change itself – if it helps resolve socially consequential problems or replace them with better ones, if it is sufficiently inclusive of different perspectives and populations, if it is more empowering than socially controlling or divisive, etc. – also impact general openness to the technology itself.

The conversation can actually advance beyond more typical binary framings (e.g. 'AI for good', 'AI is bad') to a more qualified account on when/how it might be possible for modern societies to meaningfully evolve with DMLAs without presuming that they are a priori bad or good. But technologies are also often dual-use (Brundage et al., 2018). Especially when the stakes are high, some actors might try to 'game it' to their advantage, including for adversarial purposes. Thus, an **important general consideration for international governance of DMLAs revolves around incentives for and vulnerability to abuse**.

While we cannot expect that the governance of DMLAs will be globally consensual or rule out every abuse, at a time when it seems like the digital space resembles a race, **engineering digital trust remains a critical challenge, and one particularly relevant for DMLAs.**

General perceptions around algorithmic opacity, difficulties with explainability, concerns about bias, and lack of clarity around defined responsibilities and liabilities may all, in various ways and degrees, impact the question of trust in DMLAs. That algorithms themselves cannot reflect on who bears what consequences from their use or to whom the benefits accrue; that they cannot reflect on fairness, only incorporate some mathematically defined notion of it; that they cannot realize when they are biased or evolving in paradoxical ways, how they are used in a larger system or when they are under attack are all non-negligible may all impact trust in a general sense.

However, it is also possible that what appears as a question of (mis)trust vis-à-vis a technology can be dealt with measures to improve fairness, access, etc. rather than be described as a general concern about the technology itself. While it is not easy to encapsulate what constitutes and undoes trust, two general remarks are worth noting: first, trust is not binary per se (e.g. it can vary with time and there can be different levels of trust vis-à-vis different technologies, depending on context of use, time, etc.); and second, trust cannot be mandated, but trustworthiness or trustworthy behaviour can be expected of institutions and/or people.

Insofar as 'many hands' and actors become part of the development and deployment of algorithms – from sourcing and curating the input data to processing and interpreting the outcome for a specific purpose – trustworthiness may also prove challenging to display because of such assemblage of and interlinkages between actors. **The challenge, in this respect, is about trusting the broader ecosystem** – the networks of DMLAs, the 'hands' that intervened, the motives and purposes at play, the checks and balances in both technical and non-technical sense – in which algorithmic decision-making is embedded.

Participants further elaborated on different technical tools and techniques through which governance of DMLAs can further evolve. Because algorithms can encode specific values and involve rules, it is relevant to consider which technologies and tools can embed trustworthiness.

5. Possible or emerging tools and techniques to embed trustworthiness in DMLAs

It is worthwhile recalling certain technical developments pertinent to the governance of algorithms. Of particular interest are accountable computing, blockchains, smart contracts, software verification, cryptography, and more trusted hardware technologies that can, for example, enable distributed or decentralised enforcement of accountability and transparency. Developing provable theorems that algorithms do what they are supposed to do[28] are both possible and important: they help set certain critical 'guard-rails' or ensure against classes of 'bad decision events'.

Open source processes - at the level of software and/or datasets – are important, but no panacea: open source datasets require major privacy caveats and open source software could be gamed, especially when the stakes or incentives are high. **These reservations notwithstanding, open source software remains important for developing trust in the underlying technology. It is more difficult to know if/when closed software or unverifiable play with data can be trusted**. While anonymisation, differential privacy, and privacy-preserving aggregation and analysis technologies can help produce datasets that have more transparency even when they contain sensitive data, the development of formal verification technologies for algorithms, stress tests and/or independent audits of software are crucial for assuring confidence in engineering best practices. The conversation around governance of DMLAs sometimes bifurcates on whether to regulate 'data' or 'code', but it is also possible to develop data-analysis platforms[29] that manage datasets together with code in a reproducible, accountable, privacy- and access-controlled fashion.

At a time where the technologies underlying trustworthy DMLAs remain very cutting-edge, and where the skilled workforce to develop them is extremely scarce, it may be that access to it will remain limited, for now, to well-resourced companies or organisations. This may enable these organisations to shape decision-making across the board without proper governance. Even so, technical solutions will not alone solve the problems of algorithmic bias, privacy vs. transparency, reliability/accuracy, accountability, etc. They are, however, fundamental tools to aid the governance of DMLAs, if they can be used in pursuance of specific and well-defined goals[30]. An important part of governance by technology will thus be **to define desired policy goals and general good principles in a form that can be applied to specific technologies.**

6. Concluding remarks

There is increasingly a line of resources, researchers, companies and governments – at times resembling a competitive race – working towards amassing data, 'data science' expertise and exploring which algorithmic techniques can be developed for what purposes. This eco-system is dynamic and evolving. Surveying by way of research what concrete use-cases and transversal solutions across operators, domains, etc. emerge and scale up, who is developing them, where and how, what risks and benefits actualise and to whom, etc. – remains important. Working closely with the G7, the OECD is scoping **distinct sets of principles** around artificial intelligence, such as i) general principles around democracy, human rights, privacy, etc.; ii) principles targeting the research/scientific community as well as iii)

[28] Or 'cryptographic commitments' that disclose a commitment or what an algorithm is supposed to do without revealing the internal contents of the algorithm/code itself (Kroll et al., 2016)

[29] E.g. SDSC Renga platform, EPFL (Bouillet, 2017)

[30] E.g. providing theoretical guarantees or proofs that a system is designed in such a way that certain events cannot happen or can be known if they happen, screening for privacy, regulating access, allowing 3rd party audits or testing to allow discovery of biases, detect errors, certify the data's veracity, etc.

principles focusing on government policy on issues touching jobs, skills, investment, etc. Workshop participants further suggested ongoing interdisciplinary conversations on governance of DMLAs between at least three types of representatives: domain experts for a particular use of DMLAs, experts with a crosscutting understanding of their risks, and representatives from the broader policy-making community. The sentiment, echoed throughout the workshop, is **to develop and define good standards that industry and organisations can embed in their DMLA-enabled systems and/or products 'by design'**.

If we are currently living a so-called 'hype' whereby we underestimate long-term risks and benefits and over-estimate short-term ones, **the momentum underway does not necessarily imply that we need to frame the governance of DMLAs as a one-time decision (for example a moratorium on their use)**. It rather suggests the need for **a more adaptive governance approach for a technology and ecosystem in the making.** However, when in the face of potentially irreversible damage, adaptive approaches should be very carefully crafted, in order that humans can remain in control, able to stop or control a process if it produces outcomes that are not socially desirable.

Appendix: Questions discussed in the workshop

1. The need for governance of decision-making algorithms
   - What are the key benefits and opportunities with respect to DMAs?
   - Why do we need to think about the governance of DMAs?
   - What are the immediate and long-term governance deficits with respect to DMAs?
   - What are the relevant governance possibilities?

2. Algorithmic biases, human rights and fairness
   - How can DMAs propagate biases leading to decisions that are not accurate, legally correct and/or respectful of human rights? How can biases be detected?
   - What is the relation between algorithmic and human biases? Can the former correct or better account for the latter?
   - How to design, train and use more fair DMAs?
   - Do legal norms or doctrinal practice need to change to accommodate more fair DMAs?

3. How reliable and accountable can and should DMAs be
   - What are the most likely and worrisome inaccuracies or failures?
   - Insofar as full transparency proves challenging for both technical and non-technical reasons, how to pursue more accountable analytics? What kind and degree of explainability/interpretability is technically feasible, legally possible and socially desirable?
   - Which accountability notions, goals or properties can we embed in DMAs?

4. Data protection and liability issues
   - Does the EU GDPR apply to DMAs?
   - Who is responsible if something goes wrong? How to handle the 'many hands' partaking in the design, operation and use of DMAs?
   - Where and how to specify the relevant liabilities? What, if any, changes with respect to existing legislation?

5. Trust-building measures
   - Can trustworthiness be mandated in the case of DMAs? What are the 'weakest' points in the chain of trust?
   - What concrete trust-building measures or tools at the level of technology, standards, regulation, social norms and/or international cooperation could guard against ill uses of DMAs?
   - Are new oversight bodies needed?

6. Closing the circle on the governance of DMAs
   - What approaches, concrete measures or talking points to prioritise with respect to the technical design and the governance of decision-making algorithms?
   - What to emphasise as more relevant in the long-term?

REFERENCES

This report was informed by the following literature:

Águila-Obra, A. R. del, Padilla-Meléndez, A., & Serarols-Tarrés, C. (2007). Value creation and new intermediaries on Internet. An exploratory analysis of the online news industry and the web content aggregators. *International Journal of Information Management*, *27*(3), 187–199. https://doi.org/10.1016/j.ijinfomgt.2006.12.003

Angulo, I., (2018, March 17). Facebook and YouTube should have learned from Microsoft's racist chatbot. *CNBC*. Retrieved 19 September 2018, from https://www.cnbc.com/2018/03/17/facebook-and-youtube-should-learn-from-microsoft-tay-racist-chatbot.html

Angwin, J. & Larson, J. (2016, December 30). Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. *ProPublica.* Retrieved 19 September 2018, from https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

Arai, N., (2017, April). Can a robot pass a university entrance exam? [Video file] Retrieved 19 September 2018, from https://www.ted.com/talks/noriko_arai_can_a_robot_pass_a_university_entrance_exam?language=en

Bevir, M. (2009). *Key Concepts in Governance*. SAGE.

Booker, B. (2018, August 19) HUD Hits Facebook For Allowing Housing Discrimination. Retrieved 19 September 2018, from https://www.npr.org/2018/08/19/640002304/hud-hits-facebook-for-allowing-housing-discrimination

Bouillet E., (2017). Privacy-conscious open science: A use case in biomedical research. *SDSC*. Retrieved 19 September 2018, from https://datascience.ch/privacy-conscious-open-science-a-use-case-in-biomedical-research/

Bresnahan, T. (2010). Chapter 18 - General Purpose Technologies. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the Economics of Innovation* (Vol. 2, pp. 761–791). North-Holland. https://doi.org/10.1016/S0169-7218(10)02002-2

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv:1802.07228 [Cs]*. Retrieved from http://arxiv.org/abs/1802.07228

Carpenter, J., (2015, July 6). Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/?noredirect=on&utm_term=.4fd746684ee0

Corbett-Davies, S., Pierson E., Feller A., Goel S. (2016, October 17). A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear. Retrieved 19 September 2018, from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy Non-Discrimination in Data-Driven Systems. *eprint arXiv:1707.08120*. Retrieved from http://arxiv.org/abs/1707.08120

Doneda, D., & Almeida, V. A. (2016, August). What Is Algorithm Governance? *IEEE Internet Computing*, *20*(4), 60–63.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., … Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *eprint arXiv:1711.01134*. Retrieved from http://arxiv.org/abs/1711.01134

Fumo, D. (2017, June 15). Types of Machine Learning Algorithms You Should Know. *Towards Data Science*

Greco, L., (2017, June 27). A Reality Check for IBM's Al Ambitions. *MIT Technology Review.* Retrieved 19 September 2018, from https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/

Greenwald, H. S., & Oertel, C. K. (2017). Future Directions in Machine Learning. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00079

How Policymakers Can Foster Algorithmic Accountability. (2018, May 21). Retrieved 19 September 2018, from https://www.datainnovation.org/2018/05/how-policymakers-can-foster-algorithmic-accountability/

Hvistendahl, M. (2017, December 14). In China, a Three-Digit Score Could Dictate Your Place in Society. *Wired*. Retrieved 19 September 2018, from https://www.wired.com/story/age-of-social-credit/

Klabbers, J. (1996). The Redundancy of Soft Law. *Nordic Journal of International Law,* Vol 65:2 (pp.167-182) https://doi.org/ 10.1163/15718109620294889

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). *Human Decisions and Machine Predictions*, NBER Working Paper Number 23180

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). *Accountable Algorithms* (SSRN Scholarly Paper No. ID 2765268). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2765268

Lipton, A. Z. C. (2016). The Foundations of Algorithmic Bias. Retrieved 19 September 2018, from http://approximatelycorrect.com/2016/11/07/the-foundations-of-algorithmic-bias/

Mager, A. (2012). Algorithmic ideology. *Information, Communication & Society*, *15*(5), 769–787. https://doi.org/10.1080/1369118X.2012.676056

Malgieri, G., & Comandé, G. (2017). Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law*, *7*(4), 243–265. https://doi.org/10.1093/idpl/ipx019

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Nye, J. S. (2009). Get Smart: Combining Hard and Soft Power. *Foreign Affairs*, *88*(4), 160–163.

NTSB (2018, May 24). Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy,* Crown Publishers, New York.

On-Road Automated Driving (ORAD) committee. (s. d.). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International. https://doi.org/10.4271/J3016_201609

Reilly, C., (2018, May 13). Facial-recognition software inaccurate in 98% of cases, report finds. *cnet*. Retrieved from https://www.cnet.com/news/facial-recognition-software-inaccurate-in-98-of-metropolitan-police-cases-reports/

Saurwein, F., Just, N., & Latzer, M. (2015). *Governance of Algorithms: Options and Limitations* (SSRN Scholarly Paper No. ID 2710400). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2710400

Shaffer, G. C., & Pollack, M. A. (2009). Hard vs. Soft Law: Alternatives, Complements, and Antagonists in International Governance. *Minnesota Law Review*, *94*, 706.

Smith, A. (2018, August 30). Franken-algorithms: the deadly consequences of unpredictable code. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger

Stahl, B. C. (2013). Responsible research and innovation: The role of privacy in an emerging framework. *Science and Public Policy*, *40*(6), 708–716. https://doi.org/10.1093/scipol/sct067

The battle for digital supremacy - America v China. (2018, March 15). Retrieved 19 September 2018, from https://www.economist.com/leaders/2018/03/15/the-battle-for-digital-supremacy

The world's most valuable resource is no longer oil, but data. (2017, May 6). *The Economist*. Retrieved 19 September 2018, from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Thompson, N. (2018, March 31). Emmanuel Macron Talks to WIRED About France's AI Strategy. *Wired*. Retrieved from https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/

Toyota Unintended Acceleration and the Big Bowl of "Spaghetti" Code. (2013, November 7). *Safety Research & Strategies, inc.* Retrieved September 19, 2018 from http://www.safetyresearch.net/blog/articles/toyota-unintended-acceleration-and-big-bowl-“spaghetti”-code

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., … Lin, H. (2017). FairTest: Discovering Unwarranted Associations in Data-Driven Applications (pp. 401–416). https://doi.org/10.1109/EuroSP.2017.29

Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Colorado Technology Law Journal*, *13*, 203.

Valentino-DeVries, J., Singer-Vine, J., & Soltani, A. (2012, December 24). Websites Vary Prices, Deals Based on Users' Information. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wagner, B. (2016). Algorithmic regulation and the global default: Shifting norms in Internet technology. *Etikk i Praksis - Nordic Journal of Applied Ethics*, (1), 5–13. https://doi.org/10.5324/eip.v10i1.1961

Zweig, K.A., Wenzelburger, G. & Krafft, T.D. Eur J Secur Res (2018) 3: 181. https://doi.org/10.1007/s41125-018-0031-2

ACKNOWLEDGEMENTS

IRGC thanks the following individuals who were present and contributed to the IRGC expert workshop on the governance of decision-making algorithms, 9-10 July 2018: **Kujtesë Bejtullahu Michalopoulos**, International Risk Governance Center, EPFL; **Courtney Bowman**, Palantir Technologies; **Corina Braem**, University of Zurich; **Maya Bundt,** Swiss Re**; Anne Carblanc**, OECD; **Elisa Celis**, EPFL; **Raja Chatila**, Sorbonne University & IEEE; **Pam Dixon**, World Privacy Forum; **Gilles Dubochet,** C4DT, EPFL; **Rainer Egloff,** Swiss Re; **Marie-Valentine Florin,** International Risk Governance Center, EPFL; **Bryan Ford,** EPFL; **Gero Gunkel,** Zurich Insurance; **Jerry Gupta**, SwissRe; **Hoda Heidari,** ETH Zurich; **Chris James,** HSBC & Society for Computers and Law; **James Larus,** EPFL; **Michael Latzer**, University of Zurich; **Igor Linkov,** US Army Engineer Research and Development Center & Carnegie Mellon University; **Jessica Montgomery,** The Royal Society; **Karine Perset**, OECD; **Marc Radice**, Zurich Insurance; **Jean-Marc Rickli**, Geneva Centre for Security Policy; **Elettra Ronchi,** OECD; **Anca Rusu,** International Risk Governance Center, EPFL; **Roland Scholz,** ETH Zurich; **Stephan Schreckenberg**, Swiss Re; **Eric Schuh**, Swiss Re; **Jason Stanley**, Element AI; **Rolf Staub,** Zurich Insurance; **Thomas Szucs,** Helsana Group & University of Basel; **Effy Vayena,** ETH Zurich; **Nisheeth Vishnoi**, EPFL; **Jonathan Wiene**r, Duke University.

IRGC also wishes to acknowledge and thank the Swiss Re Institute for providing a financial contribution towards the cost of the workshop and project, and the Center for Digital Trust (C4DT) at EPFL for its involvement in the preparation, facilitation and follow-up of the workshop.

This workshop report was written by Kujtese Bejtullahu Michalopoulos, with contributions and editing by Marie-Valentine Florin and Gilles Dubochet. Valuable comments were also received from other workshop participants and external experts, who helped improve the quality of this report: Maya Bund, Swiss Re, Rainer Egloff, Swiss Re; Ender Konukoglu, ETH Zurich; James Larus, EPFL; Aitana Lebrand, Swiss Institute of Bioinformatics; Granger Morgan, Carnegie Mellon University and IRGC Foundation Board; Verena Pfeiffer, University of Bern; Bernd Stahl, De Montfort University; and Philip van Hoevell, PwC.

The views and recommendations in this report do not necessarily represent the views of individual workshop participants or their employers.

ABOUT IRGC

Since 2016, IRGC consists of two distinct and independent entities, which collaborate and support each other:

- The International Risk Governance Center (IRGC@EPFL), a transdisciplinary centre at the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. More information on irgc.epfl.ch.
- The International Risk Governance Council Foundation, established in 2003 at the initiative of the Swiss government, based at EPFL, and with network partners in Europe, the US and Asia. More information on irgc.org.

These two bodies work closely together towards improving the governance of risk issues marked by complexity, uncertainty and ambiguity. A neutral platform for dialogue about emerging risks as well as opportunities and risks related to new technologies, they help improve the understanding and management of risks and opportunities by providing insight into emerging and systemic risks that have impacts on human health and safety, on the environment, on the economy and on society at large.

international risk
governance center

EPFL International Risk Governance Center

EPFL IRGC
Station 5 BAC
1015 Lausanne
Switzerland

+41 21 693 82 90
irgc@epfl.ch
irgc.epfl.ch
irgc.org

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
INTERNATIONAL RISK GOVERNANCE CENTER