Governing risk from decision-making learning algorithms (DMLAs)

Outcome from an IRGC workshop, July 2018

https://irgc.epfl.ch

November 2018

Algorithms can now learn, self-evolve and decide autonomously

Decision-making learning algorithms (DMLAs) can be understood as *information systems that use data and advanced computational techniques, including machine learning or deep learning (neural networks), to issue guidance or recommend a course of action for human actors* and/or *produce specific commands for automated systems*

- **For the time being, there is limited implementation of DMLAs**, besides a handful of industry innovators and dominant players (e.g. tech giants and certain governments). Most organisations are still *exploring* what is possible, to the extent that they are exploring the full potential that such algorithms learn, self-evolve and can make decisions autonomously.
- **The potential of DMLAs is recognized in key sectors** – notably in healthcare and for automated driving. More broadly, this huge technological revolution can also involve a profound transformation of society and the economy.

Applications

Societies are becoming increasingly dependent on digital technologies, including machine learning applied across a broad spectrum of areas such as:

- Transportation (e.g. autonomous driving)
- Health (e.g. diagnostics and prognostics, data-driven precision / genomic medicine)
- Administration (e.g. predictive policing, criminal risk assessment)
- Surveillance (e.g. citizen scoring schemes, counter-terrorism)
- Insurance (e.g. insurance underwriting, claim processing, insurance fraud detection, etc.)
- News and social media
- Advertising

Evaluating risks and opportunities from DMLAs

- Policymakers face a difficult balancing act between allowing and incentivising the meaningful uses of DMLAs from the adverse ones. Risk of wrong or unfair outcome, including possible discrimination, must be carefully evaluated in light of expected benefits in efficiency and accuracy.
- The more automated or ‘independently’ deciding algorithms are, the more they need to be scrutinized. DMLAs remain particularly challenging to decision-making when the stakes are high, when human judgment matters to concerns such as privacy, non-discrimination and confidentiality, especially when there is a risk of irreversible damage.
- Technical and governance issues are tightly interconnected. There are opportunities and risks at both levels.

Examples	*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)

DMLAs can bring many benefits to society

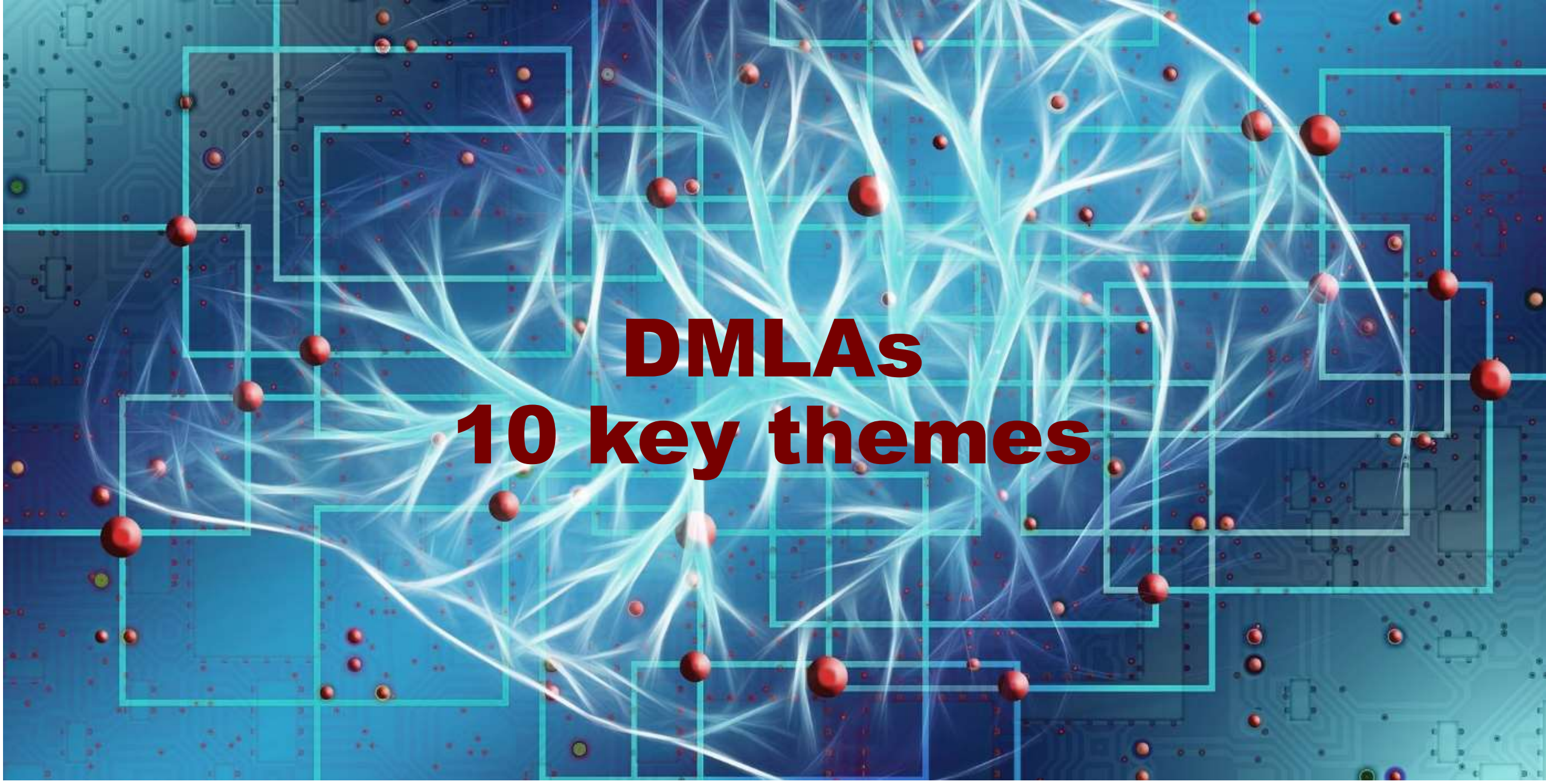
Analytic prowess	• analysing large volumes and flows of data, from multiple sources, in ways not possible for humans
Efficiency gains	• generating outcomes more promptly and less costly than could be done by human processors
Scalability	• drawing linkages, finding patterns and yielding outcomes across domains
Consistency	• processing information more consistently and systematically than humans
Adaptability	• processing and learning with dynamic data and adapting to changing inputs or variables fast
Convenience	• performing fastidious or time-consuming tasks so as to free up human time for other meaningful pursuits

DMLAs can cause new risks or amplify existing risks (1)

Erroneous or inaccurate outcomes	• difficulty to identify or correct errors or inaccuracy due to the intrinsic biases in input data and lack of transparency on the provenance of decisions, and difficulty to test DMLAs
Recurring problem of software correctness	• 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	• tension between privacy protecting rights such as 'the right to be forgotten' and the need for more complete and unbiased datasets for DMLAs to live up to their potential
Social discrimination and unfairness	• notably through the reproduction of certain undue biases around race, gender, ethnicity, age, income, etc.
Loss of accountability	• some DMLAs resemble 'black boxes' such that decision-making is difficult to understand and/or explain and thus attribution of responsibility or liability may be difficult

DMLAs can cause new risks or amplify existing risks (2)

Loss of human oversight	• DMLAs are increasingly deployed in domains (e.g. of medicine, criminal justice, etc.) where human judgment and oversight matter
Excessive surveillance and social control	• DMLAs are deployed by powerful actors, be they governments, businesses or other non-state actors 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)


DMLAs
10 key themes

#1 - Technology and governance are tightly connected

- 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 machine learning and data scientists and developers to embed the appropriate governance rules, norms or regulations into the very functioning of the algorithms.
- It is further valuable to include a mechanism of auditing and quality control, to check adherence to these rules or norms.

#2 - What is new: algorithms 'learn' and self-evolve

- Amidst different types of algorithms used for machine learning, algorithms that *learn* and *self-evolve* warrant particular attention.
- In deep learning (with e.g. neural networks) algorithms are no longer “programmed” but increasingly “learned” and adaptive, giving them an ability to perform tasks that were previously done by humans trained or entrusted for such purpose.

#3 – Evaluating risk, across domains and applications

- The evaluation of distinct and shared risks requires careful assessment of
  - undue biases in input data
  - methodological inadequacies or shortcuts caused by low-quality input data or inappropriate learning environment
  - wrong outcome, e.g. possibly resulting in social discrimination or unfairness
  - loss of accountability and of human oversight
  - inappropriate or illegal surveillance, and malignant manipulation

#4 – Governing risk, considering existing benchmarks and regulations

- When DMLAs are deployed in specialised domains –like medicine, insurance, public administration– they do not develop in a contextual vacuum: there already are certain decision-making practices, analytical thresholds, prescriptive or historical norms in place, which matter for calibrating and evaluating the performance of DMLAs vis-à-vis alternatives.
- Existing regulatory frameworks vary by domain, therefore specific applications of DMLAs require spelling out the relevant benchmarks against which their performance must be evaluated and calibrated.
- An overarching question is how to evaluate decisions by DMLAs in contrast to decisions by humans, which are not error or bias-free. When the benchmarks are lacking, how to define them?

#5 - Accuracy of outcome: critically important, not yet granted

- The accuracy (or correctness) of DMLAs' outputs is what needs to be established, especially when the decision-making process is not transparent or is difficult to explain, and/or the outcomes are difficult to interpret or explain.
- Greater attention is needed to assure more robust methodological practices (particularly as regards the quality and appropriate use of input data and learning context) and to probe the computational dynamics and learning at play.
- Of particular concern is that we do not quite know how to test machine learned and adaptive algorithms.
- DMLAs may have the potential to optimise fewer errors in aggregate, but those errors may be qualitatively worse when judged against those made for a specific individual, or in comparison to those expected from equivalent human decision-making.
- It is thus increasingly necessary to decide, perhaps even regulate for, how we determine accuracy for DMLA systems.
  - Explainability of the outcome will probably be a key component of trustworthiness
  - DMLAs should be able to detect when the outcome is inaccurate.
- Anyway, when individuals or organisations are affected by a decision which they believe is wrong or unfair, they should be given a right to receive an explanation, and a right to recourse.

#6 – Biases are a key challenge

- *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 manifesting through proxies.
- De-biasing techniques exist, but entail a trade-off: in order to evaluate whether undesired proxy measures creep into an algorithm, some very sensitive categories of information, such as race, gender, age, ethnicity, etc. may have to be included, to know if the biases are sufficiently minimised.
- Obtaining more complete and unbiased data remains a pervasive and significant challenge.
- Examples: errors in insurance contract proposal, unfairness or undue discrimination in predictive policing

#7 - Humans in control: under which circumstances and how?

- 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.
- Examples: role of the driver in automated driving, role of human judgement in criminal justice, role of medical doctor in medical diagnostic

#8 - Need to: develop standards, principles and good practices

- Standards, principles and good practices can help developers, industry, and other organisations embed best practices ‘by design’**.**
- The IEEE 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 is central

- It becomes ever more important to determine what are the concrete ways to demand and deliver accountability, who assumes what responsibility in DMLA's development and use, and who is liable in case of erroneous or wrongful applications. Better defining legal uses of DMLAs can also help different organisations determine whether to enable, accelerate or restrict their adoption.
- Regulated sectors might be able to include the use of DMLAs in their existing regulatory frames.
- Liability attribution is particularly challenging given the 'many hands' partaking in the design and deployment of DMLAs, and the lack of clarity about software liability.
- In Europe, the EU Global Data Protection Regulation (GDPR) makes in theory some provision for the right not to be subject to automated decision-making and the right to an explanation (art. 22.1). But this provision may not apply in many cases (as defined in art. 22.2), which leaves much ambiguity about which aspects of the GDPR apply to DMLAs. There remains a general prohibition on making such decisions using personal data, but interpretation of the law and compliance with it may vary by countries and/or domains.

#10 - Engineering digital and social trust is a critical challenge and increasingly relevant

- Digital trust can be ensured by techniques such as accountable computing, blockchains, smart contracts, software verification, cryptography, and 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 are both possible and important: they help set certain critical 'guard-rails' or ensure against classes of 'bad decision events'.
- While possible to look for ways to mandate or improve digital trustworthiness, the challenge is also about trusting the broader ecosystem in and around DMLAs, for which a governance structure might help. Who benefits? Whom or what to trust?
- Informational asymmetries –between data subjects, brokers, companies or various platforms where data is gathered– may affect public perception as to whether DMLAs are being put to good general use.
- 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.

Conclusion

- The development of DMLAs provides many opportunities, but also a series of technical and governance challenges around accuracy, explainability and fairness of the outcome, and transparency of the methodology and process.
- While more sharing of higher quality data is needed, data privacy must also be ensured. Social norms are changing and principles are being redefined.
- Incentivizing the production of accurate, fair and socially acceptable outcome from DMLAs will be key for the development of the technology. This will require a dialogue between scientists and society, facilitated by trusted bodies.
- Assigning accountability and legal responsibility should ensure that DMLAs will be developed for good.

https://irgc.epfl.ch

- This presentation summarises some of the findings and recommendations presented in a report elaborated by IRGC, after a multi-stakeholder and multi-disciplinary workshop on governing decision-making algorithms, on 9-10 July 2018 at the Swiss Re Institute in Zurich.
- Autorisation to reproduce granted under the condition of full acknowledgement of IRGC as the source:
  EPFL IRGC (2018) The Governance of Decision-Making Algorithms. Lausanne: EPFL International Risk Governance Center
- Available from
  - https://irgc.epfl.ch/issues/projects-cybersecurity/the-governance-of-decision-making-algorithms/
  - https://infoscience.epfl.ch/record/261264?ln=fr&p=irgc