Artificial intelligence and society

Executive summary and recommendations

Executive summary and recommendations
Artificial intelligence (AI) is one of the technologies that is transforming our society and many aspects of our daily lives. AI has already provided many positive benefits and may be a source of considerable economic prosperity. It also gives rise to questions about employment, confidentiality of data, privacy, infringement of ethical values and trust in results.

Policy makers should encourage and scientists should commit to:

- **Careful stewardship is necessary to help share the benefits of AI across society**. This will require close attention to the impact of AI on employment which will be in turn shaped by a range of factors including political, economic, and cultural elements, as well as progress in AI technologies.

- **AI systems and data should be trustworthy**. This should be facilitated through measures addressing the quality, lack of bias and traceability of data. While this can be further aided by making the data more accessible, personal data should not be made available to unauthorized parties.

- **AI systems and data should be safe and secure**. This is essential in the case of applications that involve human vulnerability and may require provably correct systems.

- **Further research is needed to help develop explainable AI systems**. When important decisions are suggested by AI impacting people, those concerned should be given sufficient information and be allowed to challenge the decisions (e.g., refuse a treatment or appeal a decision).

- **Insights from many fields are needed in order to maximize the societal benefits of AI**. Interdisciplinary research should involve diverse fields such as natural, life and medical sciences, engineering, robotics, humanities, economic and social sciences, ethics, computer science and AI itself.

- **Citizens need to be AI-ready**. A range of AI educational opportunities and information should be available and a well-founded dialogue with citizens is required to demystify this field.

- **Public policy debate on the destructive/military usage of AI should be promoted**. International undertakings limiting the risks of autonomous weapons should be considered by the relevant UN body.

- **Talent exchanges and cooperation between public research and private sector should be encouraged**. This would facilitate safe and rapid deployment of applications in areas of great human benefit. Collaboration is important for large-scale collection of data that are crucial for developing AI systems.

Introduction

AI refers to a set of methods and technologies aimed at making computers or other devices function intelligently. AI is basically a collection of algorithms operating on (usually big) data. Machine Learning (ML) is a subset of AI which deals with algorithms extracting useful information from complex data. ML applications have recently made an unanticipated impact in many areas of science and technology. There is broad consensus that AI research is progressing steadily, and that its impact on society will likely increase in the future.

The development of sophisticated algorithmic systems, combined with the availability of data and processing power, has yielded remarkable successes in various specialized tasks such as speech recognition, image classification, fault detection, autonomous vehicles, decision support systems, robotics, machine translation, legged locomotion, and question-answering systems. Some of these applications are providing extremely valuable support tools for people with disabilities. Using brain-machine interfaces, paralyzed people can interact with their environment through a computer.

Within natural and social sciences, machine learning algorithms are enabling progress and providing new tools for handling and modeling of complex data and processes, with huge potential benefits. Since a large part of what civilization has to offer is a product of human intelligence, we can only imagine what might be achieved when this intelligence is magnified by the tools AI may provide.

There is however a number of questions and concerns about potential pitfalls that require further consideration.

Progress in AI research makes it timely to focus efforts not only on making AI more capable, but also on maximizing its societal benefits while respecting ethical values. The deployment and technical developments of AI should therefore be guided by ethical considerations. Concerns are rising that biases may result from AI systems relying on statistical data analysis and machine learning.

In this general context, we first address the problems posed by AI's transformative economic impact. Second, we address the general properties that AI systems should have in order to interact satisfactorily and ethically with humans. We then address more specific issues related to the use of AI systems in healthcare, questions raised by possible AI applications to autonomous weapon systems, and consider the potential of AI embedded in robotics systems. This analysis gives rise to a set of recommendations gathered in the executive summary.

1. Managing and optimizing AI's impact on our Societies

There is a general agreement between economists and computer scientists that research needs to be done in order to maximize the economic benefits of AI while mitigating adverse effects. At this stage, it is important to consider the possible impact of AI in terms of increased inequality, unemployment and unethical behaviors. These outstanding issues are examined in further detail in what follows.

1.1 Labor market forecasting

AI could bring about significant economic benefits: across sectors, AI technologies offer the promise of boosting productivity and creating new products and services. This potential raises questions about the impact of AI on employment and working life.

AI will likely have a considerable disruptive effect on work, with certain jobs being lost, others being created, and others changing. Studies that make projections about the impact of AI on employment have high degrees of uncertainty about the rate of change, and the proportion of tasks or jobs that might be likely to be automated.

In the longer-term, technologies contribute to increased population-level productivity and wealth. However, these benefits can take time to emerge, and there can be periods in the interim where parts of the population experience dis-benefits. This suggests there may be significant transitional effects causing disruption for some people or places, and potentially widening societal inequalities in the short term. There is clearly a need for research anticipating the economic and societal impact of such

disparity, taking into account vulnerability of jobs to automation. It will be easier to analyze the impact of AI systems on various kinds of jobs, those requiring lower skilled workers and those needing highly trained professionals, than to predict the jobs that may be created in the future under various policies. There are a number of plausible future paths along which AI technologies might develop. A range of factors will play a role in shaping the impact of AI on employment, including political, economic, and cultural elements, as well as the capabilities of AI technologies. Using the best available research evidence from across disciplines can help develop policies that share across society the benefits of these technology-enabled changes.

1.2 Policies for managing and integrating AI development in society

AI will have an important impact on a range of sectors in society, augmenting or replacing human work. The challenge is to anticipate these changes and develop policies that will limit negative effects and allow a better integration of AI. Education is key both in driving AI adoption and in combating inequality.

Basic understanding of the use of data and AI technologies is needed across all ages, not only of producers and professional users of AI but for all citizens. Introducing key concepts in schools can help ensure this. Adopting a broad and balanced curriculum for educating young people in sciences, mathematics, computing, arts and humanities could equip them with a range of skills and provide a stronger basis for lifelong learning.

There is also high-demand for highly skilled employees. A range of sectors and professions will require skills to use AI in ways that are useful for them. New initiatives can help create a pool of informed users of AI systems. Support for novel apprenticeship tracks and infrastructures is also needed to build advanced skills in AI that will allow new applications with the creation of many new jobs.

These issues were already part of the declaration of Ottawa on "Realizing our digital future and shaping its impact on knowledge, industry, and the workforce" at the last G7 summit. Governments are encouraged to implement policies that will be inclusive and able to provide every citizen with equitable access to the AI benefits. This requires that information quality, security and resilience are also guaranteed as well as transparency, openness and interoperability of the AI systems.
In those areas where AI's capabilities have outpaced current regulations, there may be a need for new governance approaches that take into account ethical questions arising from human interaction with intelligent machines. It is worth emphasizing the role of humanities and social sciences broadly and in partnership with developers and users in exploring the ways in which AI may challenge existing ethical norms or indeed reveal the ways in which AI presents new ethical challenges.

2. Features of AI systems that should be encouraged

2.1 Data

Our ability to take full advantage of the synergy between AI and big data will depend in part on our capacity to acquire, critically assess and manage data. Much of the current AI technology requires access to huge volumes of data. To take full advantage of the technology, new frameworks may be necessary to make data available. This is notably true for open data and for private data of public interest where new standards might be necessary to help ensure that data can be used effectively. For example, an effort will be needed to make the meaning of data explicit, together with a representation of the context in which they have been derived, and the information about their origin and their processing. All these issues can be addressed by AI techniques, that can thus be important for keeping the many promises of open data, and providing interoperability between different types, e.g., social, economic, organizational, and technical.

At the same time, access to high quality datasets should respect privacy and confidentiality of personal data and address concerns about unfair-biases and individual rights. The best possible efforts should be made in order that access to confidential data by third parties like banks, insurance companies, potential employers is governed by regulations. Datasets must be protected against malicious attacks. Policies governing data collection, sharing and access should be in place not only for large companies but also for open source initiatives.

2.2 Performance and explainability

Some of the most successful and popular developments of AI - notably deep learning - suffer from low levels of explainability at present, and different AI methods support different types of explainability. In some cases, this might reduce the confidence that users can have in such tools. Certain domains consider explanations as essential: in medical applications, a diagnosis without explanation is unlikely to be acceptable. The tradeoffs between performance and explainability should be made explicit while aiming at developing more explainable models. The limitations of the implemented algorithms need to be described to allow users to understand the reasons for the decisions proposed by AI systems. Improving the explainability of AI can help ensure that the AI system does not introduce biases. Disparate impact has emerged as the predominant legal and theoretical concept used to designate unintended discrimination produced by the application of algorithms where a personal attribute (like ethnicity, social origins, gender and age) has a direct effect on the decisions made by the algorithm. AI systems used to make decisions which have a deep impact on the everyday life of people should not generate an undesirable disparate impact.

2.3 Verification and validation of on-line evolving systems

On-line evolving systems change in time based on the data they continuously encounter. It has recently become clear that an AI system can drift away from its initial state in an undesired fashion, for example with respect to gender and race. On-line evolving systems therefore require monitoring of the output to eventually detect undesired evolutions.

3. Exemplary fields of application and societal consequences

3.1 Health care applications

AI offers significant potential benefits in systems that support decision-making in health and care. Structural problems in this field can lead to diagnostic errors, possible failure of expertise and inefficient communication of information between research, engineering and clinical worlds. AI can help assess huge amounts of research publications, spot unlikely and weak correlations in huge data sets, analyze images and other data produced by the healthcare systems and develop new technologies. Because of the vital importance of improving clinical decision support systems, AI may contribute significantly to helping clinicians with a range of tools and devices for assisting, and complementing decision making regarding diagnosis and therapeutic options. The goal is improvement in interpreting observations and measurements, in producing diagnoses, and in making health care more accurate, effective and accessible. This requires careful system design, taking into account how AI can work alongside human users, the type of interpretability that might be necessary in different contexts, and the ways in which such systems can be verified and validated. It will be important that physicians and patients can be confident in such systems, and that these systems work well for diverse user groups.

Careful data governance is also necessary. Collaborations across countries to accelerate advances through AI is in the interest of citizens in all countries.

3.2 Autonomous weapons

AI opens new possibilities for military applications, particularly with regard to weapon systems with significant autonomy in the critical functions of selecting and attacking targets. Such autonomous weapons might lead to a new arms race, lower the threshold for war or become a tool for oppressors or terrorists. Some organizations call for a ban on autonomous weapons, similar to conventions in the chemical or biological weapons realm. Such a prohibition would require a precise definition of weapons and autonomy. In the absence of a ban of Lethal Autonomous Weapons Systems (LAWS), the compliance of any weapon system with International Humanitarian Law should be guaranteed. These weapons should be integrated into existing command and control structures in such a way that responsibility and legal accountability remain associated with specific human actors. There is a clear need for transparency and public discussion of issues raised in this area.

3.3 Robotics

Robots are sensing and moving machines embodying AI. The physical contact of the machine with the environment, including humans, is a challenge. Robots should be safe, reliable and secure. Until recently, robots were mainly used in manufacturing industry and confined to specific situations without

sharing space with humans. Today following the second wave of robotics development, robots can increasingly share the same space and interact with humans. While AI applications are focused on technologies that process data to derive knowledge for decision making, the ultimate goal of robotics is to create technical systems with capacities to interact with the physical world.
In addition to using machine learning algorithms, robotics faces fundamental constraints in terms of physical safety. Robotics design requires software certification and formal verification to maximize fault tolerance, reliability, and ability to survive.
Despite recent advancements, the expectations of progress often over estimate the pace of technological change.
Finally, in popular imagination, robotics and more generally AI are influenced by fantasy narratives rather than by scientific evidence. It is important to demystify and disseminate robotics and AI science by engaging in public education, discussion and debate with all citizens.

Royal Society
Canada

Chad Gaffield

Académie des sciences
France

Pierre Corvol

Deutsche Akademie der Naturforscher Leopoldina
Germany

Jörg Hacker

Accademia Nazionale dei Lincei
Italy

Giorgio Parisi

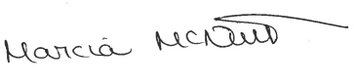
Science Council
Japan

Juichi Yamagiwa

Royal Society
United Kingdom

Venkatraman «Venki» Ramakrishnan

National Academy of Sciences
United States of America



Marcia McNutt