Lecture Notes in Artificial Intelligence 12600

Subseries of Lecture Notes in Computer Science

More information about this subseries at http://www.springer.com/series/1244

Bertrand Braunschweig ·
Malik Ghallab (Eds.)

Reflections on Artificial Intelligence for Humanity

Springer

Editors
Bertrand Braunschweig
Inria
Le Chesnay, France

Malik Ghallab
LAAS-CNRS
Toulouse, France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-69127-1 ISBN 978-3-030-69128-8 (eBook)
https://doi.org/10.1007/978-3-030-69128-8

LNCS Sublibrary: SL7 – Artificial Intelligence

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Artificial Intelligence is significantly affecting humanity. According to several thinkers and philosophers, this "soft" revolution is comparable to and as disruptive as the deployment of writing, some five thousand years ago, and printing, a few centuries ago. As media for human interaction and cognition, writing and printing have deeply changed social organizations, laws, cities, economy, and science; they have affected human values, beliefs, and religions. We are possibly witnessing a commensurately profound but much faster revolution. However, we are not just passive observers. Every person today is an actor in these dynamics, with different levels of responsibility. We all need to be well-informed, responsible actors.

We already observe the positive effects of AI in almost every field, from agriculture, industry, and services, to social interaction, knowledge dissemination, sciences, and health, including in response to pandemics. We foresee its potential to help address our sustainable development goals and the urgent challenges for the preservation of the environment.

We certainly know that there can be no human action, enterprise, or technology without risks. Those risks related to the safety, security, confidentiality, and fairness of AI systems are frequently discussed. The threats to free will of possibly manipulative systems are raising legitimate concerns. The impacts of AI on the economy, employment, human rights, equality, diversity, inclusion, and social cohesion need to be better assessed.

The ethical values to guide our choices and appraise our progress in the development and use of AI have been discussed through many initiatives, such as the principles of the Montreal declaration, the OECD principles on AI, or the EU guidelines for trustworthy AI.

The opportunities and risks are still not sufficiently well assessed. The criteria to appraise societal desirability may not be universal. Different stakeholders favor different concerns ranging from human rights and environmental preservation, to economic growth, profit, or social control. However, despite differences in deployment and views across different regions, the effects of AI will be increasingly worldwide.

The social acceptability of AI technology is not equivalent to its market acceptance. More than ensuring consumer engagement by the dissemination of convenient services at largely hidden global costs, the focus must be on social acceptability, taking into account long-term effects and possible impacts on future generations. The development and use of AI must be guided by principles of social cohesion, environmental sustainability, meaningful human activity, resource sharing, inclusion, and recognition of social and cultural differences. It has to integrate the imperatives of human rights as well as the historical, social, cultural, and ethical values of democratic societies. It needs to consider global constraints affecting the environment and international relations. It requires continued education and training as well as continual assessment of effects through social deliberation.

Research and innovation in AI are creating an avalanche of changes. These strongly depend on and are propelled by two main forces: economic competition and political initiatives. The former provides a powerful and reactive drive; however, it is mostly governed by short-term, narrow objectives. The latter rely on the former as well as on slow feedback from social awareness, education, and understanding, which strive to keep up with the pace of AI technology.

Scientists from AI and the social sciences who are involved in the progress and comprehension of the field do not have full control over its evolution, but they are not powerless; nor are they without responsibilities. They understand and guide the state of the art and what may need to be done to mitigate the negative impacts of AI. They are accountable for and capable of raising social awareness about the current limitations and risks. They can choose or at least adapt their research agenda. They can engage with integrative research and work toward socially beneficial developments. They can promote research organizations and assessment mechanisms to favor long-term, cross-disciplinary objectives addressing the social and human challenges of AI.

There is a need for a clear commitment to act in accordance with these responsibilities. Coordinated actions of all stakeholders need to be guided by the principles and values that allow us to fully assume these responsibilities, including alignment with the universal declaration of human rights, respect for and solidarity with all societies and future generations, and recognition of our interdependence with other living beings and the environment.

This book calls for all interested scientists, technologists, humanists, and concerned individuals to be involved with and to support initiatives aimed in particular at addressing the following questions[1]:

- How can we ensure the security requirements of critical applications and the safety and confidentiality of data communication and processing? What techniques and regulations for the validation, certification, and audit of AI tools are needed to develop confidence in AI? How can we identify and overcome biases in algorithms? How do we design systems that respect essential human values, ensuring moral equality and inclusion?
- What kinds of governance mechanisms are needed for personal data, metadata, and aggregated data at various levels?
- What are the effects of AI and automation on the transformation and social division of labor? What are the impacts on economic structures? What proactive and accommodation measures will be required?
- How will people benefit from decision support systems and personal digital assistants without the risk of manipulation? How do we design transparent and intelligible procedures and ensure that their functions reflect our values and criteria? How can we anticipate failure and restore human control over an AI system when it operates outside its intended scope?
- How can we devote a substantial part of our research and development resources to the major challenges of our time such as climate, environment, health, and education?

[1] Issues addressed by the Global Forum on AI for Humanity, Paris, Oct. 28–30, 2019.

The above issues raise many scientific challenges specific to AI, as well as interdisciplinary challenges for the sciences and humanities. They must be the topic of interdisciplinary research, social observatories and experiments, citizen deliberations, and political choices. They must be the focus of international collaborations and coordinated global actions.

The "Reflections on AI for Humanity" proposed in this book develop the above problems and sketch approaches for solving them. They aim at supporting the work of forthcoming initiatives in the field, in particular of the *Global Partnership on Artificial Intelligence*, a multilateral initiative launched in June 2020 by fourteen countries and the European Union. We hope that they will contribute to building a better and more responsible AI.

December 2020

Bertrand Braunschweig
Malik Ghallab

Organization

Programme Committee of the Global Forum for Artificial Intelligence for Humanity, October 28–30 2019, Paris

Pekka Ala-Pietilä	Huhtamaki, Finland
Elisabeth André	University of Augsburg, Germany
Noriko Arai	National Institute of Informatics, Japan
Genevieve Bell	Australian National University, Australia
Bertrand Braunschweig (Co-chair)	Inria, France
Natalie Cartwright	Finn AI, Canada
Carlo Casonato	University of Trento, Italy
Claude Castelluccia	Inria, France
Raja Chatila	Sorbonne University, France
Kate Crawford	AI Now Institute and Microsoft, USA
Sylvie Delacroix	University of Birmingham and Alan Turing Institute, UK
Andreas Dengel	DFKI, Germany
Laurence Devillers	Sorbonne University, France
Virginia Dignum	Umeå University, Sweden
Rebecca Finlay	CIFAR, Canada
Françoise Fogelman- Soulié	Hub France IA, France
Malik Ghallab (Co-chair)	CNRS, France
Alexandre Gefen	CNRS, France
Yuko Harayama	RIKEN, Japan
Martial Hebert	Carnegie Mellon University, USA
Holger Hoos	Universiteit Leiden, Netherlands
Lyse Langlois	Observatoire international sur les impacts sociétaux de l'intelligence artificielle et du numérique (OBVIA), Canada
Fei-Fei Li	Stanford University, USA
Jocelyn Maclure	Laval University, Canada
Ioana Manolescu	Inria and École polytechnique, France
Joel Martin	National Research Council, Canada
Michela Milano	University of Bologna, Italy
Katharina Morik	Technical University of Dortmund, Germany
Joëlle Pineau	McGill University and Facebook, Canada
Stuart Russell	University of California, Berkeley, USA
Bernhard Schölkopf	Max Planck Institute for Intelligent Systems, Germany and ETH Zurich, Switzerland
Hideaki Takeda	National Institute of Informatics, Japan

Paolo Traverso	Fondazione Bruno Kessler, Italy
Junichi Tsujii	National Institute of Advanced Industrial Science and Technology, Japan
Hyun Seung Yang	Korea Advanced Institute of Science and Technology, Korea

Contents