

Philosophical Studies Series

Luciano Floridi *Editor*

Ethics, Governance, and Policies in Artificial Intelligence



Springer

Philosophical Studies Series

Volume 144

Editor-in-Chief

Mariarosaria Taddeo, Oxford Internet Institute, University of Oxford, Oxford, UK

Editorial Board Member

Patrick Allo, Vrije Universiteit Brussel, Brussel, Belgium

Advisory Editors

Lynne Baker, Department of Philosophy, University of Massachusetts, Amherst, USA

Stewart Cohen, Arizona State University, Tempe, AZ, USA

Radu Bogdan, Department of Philosophy, Tulane University, New Orleans, LA, USA

Marian David, Karl-Franzens-Universität, Graz, Austria

John Fischer, University of California, Riverside, Riverside, CA, USA

Keith Lehrer, University Of Arizona, Tucson, AZ, USA

Denise Meyerson, Macquarie University, Sydney, Australia

Francois Recanati, Ecole Normale Supérieure, Institut Jean Nicod, Paris, France

Mark Sainsbury, University of Texas at Austin, Austin, TX, USA

Barry Smith, State University of New York at Buffalo, Buffalo, NY, USA

Nicholas Smith, Department of Philosophy, Lewis and Clark College, Portland, OR, USA

Linda Zagzebski, Department of Philosophy, University of Oklahoma, Norman, OK, USA

Philosophical Studies Series aims to provide a forum for the best current research in contemporary philosophy broadly conceived, its methodologies, and applications. Since Wilfrid Sellars and Keith Lehrer founded the series in 1974, the book series has welcomed a wide variety of different approaches, and every effort is made to maintain this pluralism, not for its own sake, but in order to represent the many fruitful and illuminating ways of addressing philosophical questions and investigating related applications and disciplines.

The book series is interested in classical topics of all branches of philosophy including, but not limited to:

- Ethics
- Epistemology
- Logic
- Philosophy of language
- Philosophy of logic
- Philosophy of mind
- Philosophy of religion
- Philosophy of science

Special attention is paid to studies that focus on:

- the interplay of empirical and philosophical viewpoints
- the implications and consequences of conceptual phenomena for research as well as for society
- philosophies of specific sciences, such as philosophy of biology, philosophy of chemistry, philosophy of computer science, philosophy of information, philosophy of neuroscience, philosophy of physics, or philosophy of technology; and
- contributions to the formal (logical, set-theoretical, mathematical, information-theoretical, decision-theoretical, etc.) methodology of sciences.

Likewise, the applications of conceptual and methodological investigations to applied sciences as well as social and technological phenomena are strongly encouraged.

Philosophical Studies Series welcomes historically informed research, but privileges philosophical theories and the discussion of contemporary issues rather than purely scholarly investigations into the history of ideas or authors. Besides monographs, *Philosophical Studies Series* publishes thematically unified anthologies, selected papers from relevant conferences, and edited volumes with a well-defined topical focus inside the aim and scope of the book series. The contributions in the volumes are expected to be focused and structurally organized in accordance with the central theme(s), and are tied together by an editorial introduction. Volumes are completed by extensive bibliographies.

The series discourages the submission of manuscripts that contain reprints of previous published material and/or manuscripts that are below 160 pages/88,000 words.

For inquiries and submission of proposals authors can contact the editor-in-chief Mariarosaria Taddeo via: mariarosaria.taddeo@oii.ox.ac.uk

More information about this series at <http://www.springer.com/series/6459>


Luciano Floridi

Editor

Ethics, Governance, and Policies in Artificial Intelligence

 Springer

Editor

Luciano Floridi 
Oxford Internet Institute
University of Oxford
Oxford, UK

ISSN 0921-8599

ISSN 2542-8349 (electronic)

Philosophical Studies Series

ISBN 978-3-030-81906-4

ISBN 978-3-030-81907-1 (eBook)

<https://doi.org/10.1007/978-3-030-81907-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Acknowledgements

I shall not repeat here the acknowledgements that can be found in each chapter and corresponding article, but rather thank all the people who have made this book possible. First of all, Danuta Farah, my personal assistant. She carefully and patiently edited the original articles and skilfully managed the production process. Without her contribution and organisational support, this book would have been impossible. Next, my colleague and co-director of the Digital Ethics Lab, Mariarosaria Taddeo, for her many ideas and suggestions in the past that led to this book, and her encouragement to pursue the project of a unified anthology of “the best of” in the ethics of AI by the DELab. And finally, all the authors whose brilliant intellectual work is showcased in this volume: Nikita Aggarwal, Monica Beltrametti, Christopher Burr, Raja Chatila, Patrice Chazerand, Josh Cows, Alexander Denev, Virginia Dignum, Anat Elhalal, Robert Gorwa, Indra Joshi, Thomas C. King, Libby Kinsey, Michelle Seng Ah Lee, Christoph Luetge, Caio C. V. Machado, Tom McCutcheon, Robert Madelin, Jessica Morley, Carl Ohman, Ugo Pagallo, Huw Roberts, Francesca Rossi, Burkhard Schafer, Mariarosaria Taddeo, Andreas Tsamados, Peggy Valcke, Effy Vayena, Vincent Wang, and David S. Watson.

Contents

1	Introduction – The Importance of an Ethics-First Approach to the Development of AI	1
	Luciano Floridi	
2	A Unified Framework of Five Principles for AI in Society	5
	Luciano Floridi and Josh Cowls	
3	An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	19
	Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena	
4	Establishing the Rules for Building Trustworthy AI	41
	Luciano Floridi	
5	The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation	47
	Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi	
6	Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical	81
	Luciano Floridi	
7	How AI Can Be a Force for Good – An Ethical Framework to Harness the Potential of AI While Keeping Humans in Control	91
	Mariarosaria Taddeo and Luciano Floridi	

8	The Ethics of Algorithms: Key Problems and Solutions	97
	Andreas Tsamados, Nikita Aggarwal, Josh Cows, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi	
9	How to Design AI for Social Good: Seven Essential Factors	125
	Luciano Floridi, Josh Cows, Thomas C. King, and Mariarosaria Taddeo	
10	From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices	153
	Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal	
11	The Explanation Game: A Formal Framework for Interpretable Machine Learning	185
	David S. Watson and Luciano Floridi	
12	Artificial Agents and Their Moral Nature	221
	Luciano Floridi	
13	Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions	251
	Thomas C. King, Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi	
14	Regulate Artificial Intelligence to Avert Cyber Arms Race	283
	Mariarosaria Taddeo and Luciano Floridi	
15	Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword	289
	Mariarosaria Taddeo, Tom McCutcheon, and Luciano Floridi	
16	Prayer-Bots and Religious Worship on Twitter: A Call for a Wider Research Agenda	299
	Carl Öhman, Robert Gorwa, and Luciano Floridi	
17	Artificial Intelligence, Deepfakes and a Future of Ectypes	307
	Luciano Floridi	
18	The Ethics of AI in Health Care: A Mapping Review	313
	Jessica Morley, Caio C. V. Machado, Christopher Burr, Josh Cows, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi	
19	Autonomous Vehicles: From Whether and When to Where and How	347
	Luciano Floridi	

20 Innovating with Confidence: Embedding AI Governance and Fairness in a Financial Services Risk Management Framework . . . 353
Michelle Seng Ah. Lee, Luciano Floridi, and Alexander Denev

21 Robots, Jobs, Taxes, and Responsibilities 373
Luciano Floridi

22 What the Near Future of Artificial Intelligence Could Be 379
Luciano Floridi

Contributors

Nikita Aggarwal Faculty of Law, Oxford Internet Institute, University of Oxford, Oxford, UK

Monica Beltrametti Naver Corporation, Grenoble, France

Christopher Burr Alan Turing Institute, London, UK

Raja Chatila French National Center of Scientific Research, Paris, France
Institute of Intelligent Systems and Robotics at Pierre, Marie Curie University, Paris, France

Patrice Chazerand Digital Europe, Brussels, Belgium

Josh Cows Oxford Internet Institute, University of Oxford, Oxford, UK
Alan Turing Institute, London, UK

Alexander Denev Deloitte, London, UK

Virginia Dignum Department of Computing Science, University of Umeå, Umeå, Sweden

Delft Design for Values Institute, Delft University of Technology, Delft, the Netherlands

Anat Elhalal Digital Catapult, London, UK

Luciano Floridi Oxford Internet Institute, University of Oxford, Oxford, UK

Robert Gorwa Department of Politics and International Relations, Saint Anthony's College, University of Oxford, Oxford, UK

Indra Joshi NHSX, London, UK

Thomas C. King Oxford Internet Institute, University of Oxford, Oxford, UK
Amherst, Cheltenham, UK

Libby Kinsey Digital Catapult, London, UK

Michelle Seng Ah. Lee University of Cambridge, Cambridge, UK

Christoph Luetge TUM School of Governance, Technical University of Munich,
Munich, Germany

Caio C. V. Machado Oxford Internet Institute, University of Oxford, Oxford, UK

Robert Madelin Defence Science and Technology Laboratories, Salisbury, UK
Centre for Technology and Global Affairs, University of Oxford, Oxford, UK

Tom McCutcheon Defence Science and Technology Laboratories, Salisbury, UK

Jessica Morley Oxford Internet Institute, University of Oxford, Oxford, UK

Carl Ohman Uppsala University, Uppsala, Sweden

Ugo Pagallo Department of Law, University of Turin, Turin, Italy

Huw Roberts Oxford Internet Institute, University of Oxford, Oxford, UK

Francesca Rossi IBM Research, Albany, NY, USA
University of Padova, Padova, Italy

Burkhard Schafer School of Law, University of Edinburgh Law School,
Edinburgh, UK

Mariarosaria Taddeo Oxford Internet Institute, University of Oxford, Oxford, UK
Alan Turing Institute, London, UK

Andreas Tsamados Oxford Internet Institute, University of Oxford, Oxford, UK

Peggy Valcke Centre for IT & IP Law, Catholic University of Leuven, Leuven,
Flanders, Belgium
Bocconi University, Milan, Italy

Effy Vayena Bioethics, Health Ethics and Policy Lab, ETH Zurich, Zurich,
Switzerland

Vincent Wang Department of Computer Science, University of Oxford, Oxford,
UK

David S. Watson Oxford Internet Institute, University of Oxford, Oxford, UK
Department of Statistical Science, University College London, London, UK

Chapter 1

Introduction – The Importance of an Ethics-First Approach to the Development of AI



Luciano Floridi 

Abstract This is the introduction to the volume. It highlights the various “seasons” through which the development of AI has gone, and how the failures and successes of AI raise ethical questions, and require an ethical approach.

Keywords Artificial Intelligence (AI) · Ethics of AI · Summer of AI · Winter of AI

The trouble with seasonal metaphors is that they are cyclical. If you say that artificial intelligence (AI) got through a bad winter, you must also remember that winter will return, and you better be ready. An AI winter is that stage when technology, business, and the media get out of their warm and comfortable bubble, cool down, temper their sci-fi speculations and unreasonable hypes, and come to terms with what AI can or cannot really do as a technology (Floridi 2019), without exaggeration. Investments become more discerning, and journalists stop writing about AI, to chase some other fashionable topics and fuel the next fad.

AI has had several winters.¹ Among the most significant, there was one in the late seventies, and another at the turn of the eighties and nineties. Today, we are talking about another predictable winter (Nield 2019; Walch 2019; Schuchmann 2019).² AI is subject to these hype cycles because it is a hope or fear that we have entertained since we were thrown out of paradise: some form of agency that does everything for us, instead of us, better than us, with all the dreamy advantages (we shall be on holiday forever) and the nightmarish risks (we are going to be enslaved) that this entails. For some people, speculating about all this is irresistible. It is the wild west of

¹https://en.wikipedia.org/wiki/AI_winter

²Even the BBC, which has contributed to the hype (see for example: <https://www.bbc.co.uk/programmes/p031wmt7>), now acknowledges it might have been... a hype: <https://www.bbc.co.uk/news/technology-51064369>

L. Floridi (✉)

Oxford Internet Institute, University of Oxford, Oxford, UK

e-mail: luciano.floridi@oii.ox.ac.uk

“what if” scenarios. But I hope the reader will forgive me for a “I told you so” moment. For some time, I have been warning against commentators and “experts”, who were competing to see who could tell the tallest tale (Floridi 2016). A web of myths ensued. They spoke of AI as if it were the ultimate panacea, which would solve everything and overcome everything; or as the final catastrophe, a superintelligence that would destroy millions of jobs, replacing lawyers and doctors, journalists and researchers, truckers and taxi drivers, and ending by dominating human beings as if they were pets at best. Many followed Elon Musk in declaring the development of AI the greatest existential risk run by humanity. As if most of humanity did not live in misery and suffering. As if wars, famine, pollution, global warming, social injustice, and fundamentalism were science fiction, or just negligible nuisances, unworthy of their considerations. They insisted that law and regulations were always going to be too late and never catch up with AI, when in fact laws and norms are not about the speed but about the direction of innovation, for they should steer the proper development of a society (if we like where we are heading, we cannot go there quickly enough). Today, we know that legislation is coming, at least in the EU. They claimed AI was a magic black box, which we could never explain, when in fact it is a matter of the correct level of abstraction (Floridi 2008) at which to interpret the complex interactions engineered – even car traffic downtown becomes a black box if you wish to know why every single individual is there at that moment. Today there is a growing development of adequate tools to monitor and understand how machine learning systems reach their outcomes (Watson and Floridi 2020). They spread scepticism about the possibility of an ethical framework that would synthesise what we mean by socially good AI, when in fact the EU, the OECD, and China have converged on very similar principles that offer a common platform for further agreements (Floridi and Cowls 2019). Sophists in search of headlines. They should be ashamed and apologize. Not only for their untenable comments, but also for the great irresponsibility and alarmism, which have misled public opinion both about a potentially useful technology – that could provide helpful solutions, from medicine to security and monitoring systems (Taddeo and Floridi 2018) – and about the real risks – which we know are concrete but so much less fancy, from everyday manipulation of choices (Milano et al. 2019) to increased pressure on individual and group privacy (Floridi 2014), from cyberconflicts to the use of AI by organised crime for money laundering and identity theft (King et al. 2020).

The risk of every AI summer is that over-inflated expectations turn into a mass distraction. The risk of every AI winter is that the backlash is excessive, the disappointment too negative, and potentially valuable solutions are thrown out with the water of the illusions. Managing the world is an increasingly complex task: megacities and their “smartification” offer a good example. And we have planetary problems – such as global warming, social injustice, and migration – which require ever higher degrees of coordination to be solved. It seems obvious that we need all the good technology that we can design, develop, and deploy to cope with these challenges, and all human intelligence we can exercise to put this technology in the service of a better future. AI can play an important role in all

this because we need increasingly smarter ways of processing immense quantities of data, sustainably and efficiently. But AI must be treated as a normal technology, neither as a miracle nor as a plague, and as one of the many solutions that human ingenuity has managed to devise. This is also why the ethical debate is and will always remain an entirely human question, and a very crucial one, as this volume shows.

Now that the new winter is coming, we may try to learn some lessons, and avoid this yo-yo of unreasonable illusions and exaggerated disillusion. Let us not forget that the winter of AI should not be the winter of its opportunities. It certainly won't be the winter of its risks and ethical challenges. We need to ask ourselves whether AI solutions are really going to *replace* previous solutions – as the automobile has done with the carriage – *diversify* them – as did the motorcycle with the bicycle – or *complement* and *expand* them – as the digital smart watch has done with the analog one. What will the level of social acceptability or preferability be in whatever way AI survives the new winter? Are we really going to be wearing some kind of strange glasses to live in a virtual or augmented world created by AI? Consider that today many people are reluctant to wear glasses even when they seriously need them, just for aesthetic reasons. And then, are there feasible AI solutions in everyday life? Are the necessary skills, datasets, infrastructure, and business models in place to make an AI application successful? The futurologists find these questions boring. They like a single, simple idea, which interprets and changes everything, that can be spread thinly across an easy book that makes the reader feel intelligent, a book to be read by everyone today and ignored by all tomorrow. It is the bad diet of junk fast-food for thoughts and the curse of the airport bestseller. We need to resist oversimplification. This time let us think more deeply and extensively on what we are doing and planning with AI. The exercise is called philosophy, not futurology.

This volume is meant to contribute to such an exercise in slower and deeper thinking. It collects some of the most significant outcomes of the research on the ethics of AI conducted by members of the Digital Ethics Lab (DELab), the OII research group that I direct at the University of Oxford, also in collaboration with other colleagues. The chapters have appeared before in a variety of peer-reviewed, international journals, but never together. For the sake of consistency, they have not been modified in content, only in format. The hope is that the reader will find having the whole collection in one place not just convenient, but also intellectually useful, to see the patterns and developments in reasonings and conclusions. As the ethical debate on AI becomes increasingly specialised, mainstream, and practically oriented, the hope is that the chapter in this book may help establish a robust foundation for further studies. Whether this hope is realistic only the reader can judge.

References

- Floridi, Luciano. 2008. The method of levels of abstraction. *Minds and Machines* 18 (3): 303–329.
- . 2014. Open data, data protection, and group privacy. *Philosophy & Technology* 27 (1): 1–3.
- . 2016. Should we be afraid of AI. *Aeon Essays*. <https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible>
- . 2019. What the near future of artificial intelligence could be. *Philosophy & Technology* 32 (1): 1–15. <https://doi.org/10.1007/s13347-019-00345-y>.
- Floridi, Luciano, and Josh Cowls. 2019. A unified framework of five principles for AI in society. *Harvard Data Science Review* 1 (1): 99.
- King, Thomas C., Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. 2020. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics* 26 (1): 89–120.
- Milano, Silvia, Mariarosaria, Taddeo, and Luciano, Floridi. 2019. *Recommender systems and their ethical challenges*. Available at SSRN 3378581.
- Nield, Thomas. 2019. Is deep learning already hitting its limitations? And is another AI winter coming? *Towards Data Science*, January 5. <https://towardsdatascience.com/is-deep-learning-already-hitting-its-limitations-c81826082ac3>
- Schuchmann, Sebastian. 2019. Probability of an approaching AI winter. *Towards Data Science*, August 17. <https://towardsdatascience.com/probability-of-an-approaching-ai-winter-c2d818fb338a>
- Taddeo, Mariarosaria, and Luciano Floridi. 2018. How AI can be a force for good. *Science* 361 (6404): 751–752.
- Walch, Kathleen. 2019. Are we heading for another AI winter soon? *Forbes*, October 20. <https://www.forbes.com/sites/cognitiveworld/2019/10/20/are-we-heading-for-another-ai-winter-soon/#783bf81256d6>
- Watson, David S., and Luciano Floridi. 2020. The explanation game: A formal framework for interpretable machine learning. *Synthese*. <https://doi.org/10.1007/s11229-020-02629-9>.

Chapter 2

A Unified Framework of Five Principles for AI in Society



Luciano Floridi  and Josh Cows 

Abstract Artificial Intelligence (AI) is already having a major impact on society. As a result, many organizations have launched a wide range of initiatives to establish ethical principles for the adoption of socially beneficial AI. Unfortunately, the sheer volume of proposed principles threatens to overwhelm and confuse. How might this problem of ‘principle proliferation’ be solved? In this paper, we report the results of a fine-grained analysis of several of the highest-profile sets of ethical principles for AI. We assess whether these principles converge upon a set of agreed-upon principles, or diverge, with significant disagreement over what constitutes ‘ethical AI.’ Our analysis finds a high degree of overlap among the sets of principles we analyze. We then identify an overarching framework consisting of five core principles for ethical AI. Four of them are core principles commonly used in bioethics: beneficence, non-maleficence, autonomy, and justice. On the basis of our comparative analysis, we argue that a new principle is needed in addition: explicability, understood as incorporating both the epistemological sense of intelligibility (as an answer to the question ‘how does it work?’) and in the ethical sense of accountability (as an answer to the question: ‘who is responsible for the way it works?’). In the ensuing discussion, we note the limitations and assess the implications of this ethical framework for future efforts to create laws, rules, technical standards, and best practices for ethical AI in a wide range of contexts.

Keywords Accountability · Autonomy · Artificial Intelligence · Beneficence · Ethics · Explicability · Fairness · Intelligibility · Justice · Non-maleficence

L. Floridi (✉)

Oxford Internet Institute, University of Oxford, Oxford, UK
e-mail: luciano.floridi@oii.ox.ac.uk

J. Cows

Oxford Internet Institute, University of Oxford, Oxford, UK

Alan Turing Institute, London, UK

e-mail: josh.cows@oii.ox.ac.uk

2.1 Introduction

Artificial Intelligence (AI) is already having a major impact on society. The key questions are how, where, when, and by whom the impact of AI will be felt. As a result, many organizations have launched a wide range of initiatives to establish ethical principles for the adoption of socially beneficial AI. Unfortunately, the sheer volume of proposed principles threatens to become overwhelming and confusing, posing two potential problems.¹ Either the various sets of ethical principles for AI are similar, leading to unnecessary repetition and redundancy, or, if they differ significantly, confusion and ambiguity will result instead. The worst outcome would be a ‘market for principles’ where stakeholders may be tempted to ‘shop’ for the most appealing ones (Floridi 2019b).

How might this problem of ‘principle proliferation’ be solved? In this paper, we report the results of a fine-grained analysis of several of the highest-profile sets of ethical principles for AI. We assess whether these principles are convergent, with a set of agreed-upon principles, or divergent, with significant disagreement over what constitutes ‘ethical AI.’ Our analysis finds a high degree of overlap among the sets of principles we analyze. We then identify an overarching framework consisting of five core principles for ethical AI. In the ensuing discussion, we note the limitations and assess the implications of this ethical framework for future efforts to create laws, rules, standards, and best practices for ethical AI in a wide range of contexts.

2.2 Artificial Intelligence: A Research Area in Search of a Definition

AI has been defined in many ways. Today, it comprises several techno-scientific branches, well summarized in Fig. 2.1 (see also the articles by Dick and Jordan in this issue for enlightening analyses).

Altogether, AI paradigms still satisfy the classic definition provided by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon in their seminal Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, the founding document and later event that established the new field of AI in 1955:

For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving. (Quotation from the 2006 re-issue in McCarthy et al. 2006 [1955]).

This is a counterfactual: were a human to behave in that way, that behaviour would be called intelligent. It does not mean that the machine is intelligent, or even thinking. The latter scenario is a fallacy, and smacks of superstition. Just because a

¹These are not the only problems, see (Floridi 2019b).

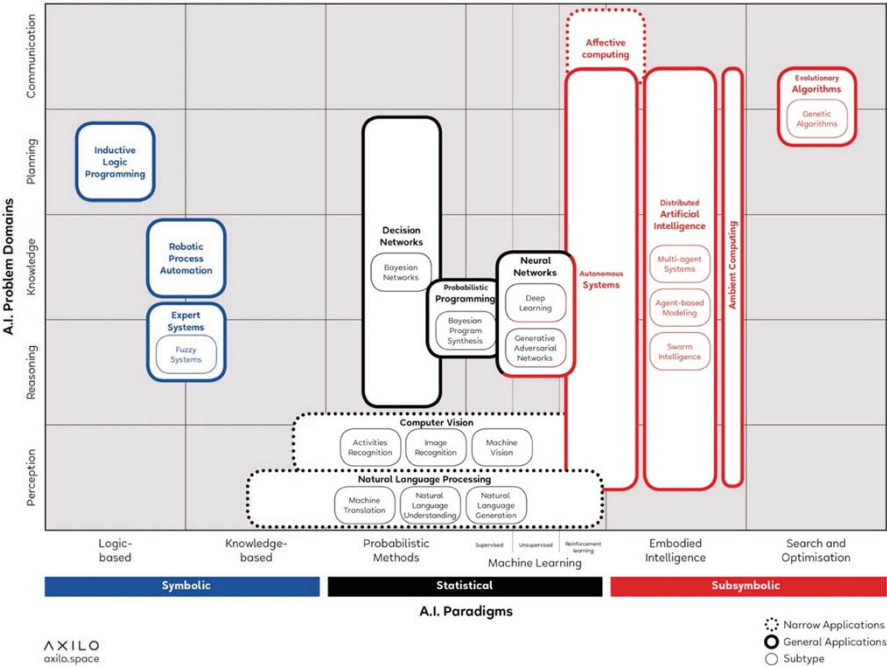


Fig. 2.1 AI Knowledge Map (AIKM). (Source: Corea (2019), reproduced with permission courtesy of F. Corea)

dishwasher cleans the dishes as well as (or even better than) I do does not mean that it cleans them like I do, or needs any intelligence to achieve its task. The same counterfactual understanding of AI underpins the Turing test (Floridi et al. 2009), which, in this case, checks the ability of a machine to perform a task in such a way that the outcome would be indistinguishable from the outcome of a human agent working to achieve the same task (Turing 1950).

The classic definition enables one to conceptualize AI as a growing resource of interactive, autonomous, and often self-learning agency (in the machine learning sense, see Fig. 2.1), that can deal with tasks that would otherwise require human intelligence and intervention to be performed successfully. In short, AI is defined on the basis of engineered outcomes and actions and so, in what follows, we shall treat AI as a reservoir of smart agency on tap (see also Floridi 2019a). This is sufficiently general to capture the many ways in which AI is discussed in the documents we analyse in the rest of this article.

2.3 A Unified Framework of Five Principles for Ethical AI

The establishment of artificial intelligence as a field of academic research dates back to the 1950s (McCarthy et al. 2006 [1955]). The ethical debate is almost as old (Samuel 1960; Wiener 1960). However, it is only in recent years that impressive advances in the capabilities and applications of AI systems have brought the opportunities and risks of AI for society into sharper focus (Yang et al. 2018). The increasing demand for reflection and clear policies on the impact of AI on society has yielded a glut of initiatives. Each additional initiative yields a supplementary statement of principles, values, or tenets to guide the development and adoption of AI. The risk is unnecessary repetition and overlap, if the various sets of principles are similar, or confusion and ambiguity, if they differ. In either eventuality, the development of laws, rules, standards, and best practices to ensure that AI is socially beneficial may be delayed by the need to navigate the wealth of principles and declarations set out by an ever-expanding array of initiatives.

The time has come for a comparative analysis of these documents, including an assessment of whether they converge or diverge and, if the former, whether a unified framework may therefore be synthesised. For this comparative analysis, we identified six high-profile initiatives established in the interest of socially beneficial AI:

1. The Asilomar AI Principles, developed under the auspices of the Future of Life Institute, in collaboration with attendees of the high-level Asilomar conference of January 2017 (hereafter ‘Asilomar’; Asilomar AI Principles 2017)
2. The Montreal Declaration for Responsible AI, developed under the auspices of the University of Montreal, following the Forum on the Socially Responsible Development of AI of November 2017 (hereafter ‘Montreal’; Montreal Declaration 2017)²
3. The General Principles offered in the second version of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. This crowd-sourced global treatise received contributions from 250 global thought leaders to develop principles and recommendations for the ethical development and design of autonomous and intelligent systems, and was published in December 2017 (hereafter ‘IEEE’; IEEE 2017, p. 6)³
4. The Ethical Principles offered in the Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems, published by the European Commission’s European Group on Ethics in Science and New Technologies, in March 2018 (hereafter ‘EGE’; EGE 2018, pp. 16–20)

²The Montreal Declaration is currently open for comments as part of a redrafting exercise. The principles we refer to here are those which were publicly announced as of May 1, 2018.

³The third version of Ethically Aligned Design will be released in 2019 following wider public consultation.

5. The ‘five overarching principles for an AI code’ offered in UK House of Lords Artificial Intelligence Committee’s report, *AI in the UK: ready, willing and able?*, published in April 2018 (hereafter ‘AIUK’; House of Lords 2018, §417)
6. The Tenets of the Partnership on AI, a multi-stakeholder organization consisting of academics, researchers, civil society organisations, companies building and utilising AI technology, and other groups (hereafter ‘the Partnership’; Partnership on AI 2018).

Each set of principles meets three basic criteria: they are recent, published within the last 3 years; directly relevant to AI and its impact on society as a whole (thus excluding documents specific to a particular domain, industry, or sector); and highly reputable, published by authoritative, multi-stakeholder organizations with at least national scope.⁴ Taken together, they yield 47 principles.⁵ Overall, we find a degree of coherence and overlap between the six sets of principles that is impressive and reassuring. This convergence can most clearly be shown by comparing the sets of principles with the four core principles commonly used in bioethics: beneficence, non-maleficence, autonomy, and justice (Beauchamp and Childress 2012). The comparison should not be surprising. Of all areas of applied ethics, bioethics is the one that most closely resembles digital ethics in dealing ecologically with new forms of agents, patients, and environments (Floridi 2013). Yet while the four bioethical principles adapt surprisingly well to the fresh ethical challenges posed by artificial intelligence, they do not offer a perfect translation. As we shall see, the underlying meaning of each of the principles is contested, with similar terms often used to mean different things. Nor are the four principles exhaustive. On the basis of our comparative analysis, we argue that a new principle is needed in addition: explicability, understood as incorporating both intelligibility (for non-experts, e.g., patients or business customers, and for experts, e.g., product designers or engineers) and accountability. However, the convergence that we detect between these different sets of principles also demands caution. We explain the reasons for this caution in the following section, but first, we introduce the five principles.

⁴ A similar evaluation of AI ethics guidelines has recently been undertaken by Hagendorff (2019), which adopts different criteria of inclusion and assessment. Note that the evaluation includes in its sample the set of principles we describe here.

⁵ Of the six documents, the Asilomar Principles offer the largest number of principles with arguably the broadest scope. The 23 principles are organised under three headings, “research issues”, “ethics and values”, and “longer-term issues”. We have omitted consideration of the five “research issues” here as they are related specifically to the practicalities of AI development in the narrower context of academia and industry. Similarly, the Partnership’s eight Tenets consist of both intra-organisational objectives and wider principles for the development and use of AI. We include only the wider principles (the first, sixth, and seventh tenets).

2.3.1 Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet

The principle of creating AI technology that is beneficial to humanity is expressed in different ways across the six documents, but is perhaps the easiest of the four traditional bioethics principles to observe. Montreal and IEEE principles both use the term “well-being”; for Montreal, “the development of AI should ultimately promote the well-being of all sentient creatures,” while IEEE states the need to “prioritize human well-being as an outcome in all system designs.” AIUK and Asilomar both characterise this principle as the “common good”: AI should “be developed for the common good and the benefit of humanity,” according to AIUK. The Partnership describes the intention to “ensure that AI technologies benefit and empower as many people as possible”, while the EGE emphasizes the principle of both “human dignity” and “sustainability.” Its principle of “sustainability” articulates perhaps the widest of all interpretations of beneficence, arguing that “AI technology must be in line with . . . ensur[ing] the basic preconditions for life on our planet, continued prospering for mankind and the preservation of a good environment for future generations.” Taken together, the prominence of beneficence firmly underlines the central importance of promoting the well-being of people and the planet with AI.

2.3.2 Non-maleficence: Privacy, Security and ‘Capability Caution’

Though ‘do only good’ (beneficence) and ‘do no harm’ (non-maleficence) may seem logically equivalent, they are not, and represent distinct principles. While the six documents all encourage the creation of beneficent AI, each one also cautions against various negative consequences of overusing or misusing AI technologies (Cows et al. 2018). Of particular concern is the prevention of infringements on personal privacy, which is included as a principle in five of the six sets. Several of the documents emphasize avoiding the misuse of AI technologies in other ways. The Asilomar Principles warn against the threats of an AI arms race and of the recursive self-improvement of AI, while the Partnership similarly asserts the importance of AI operating “within secure constraints.” The IEEE document meanwhile cites the need to “avoid misuse,” and the Montreal Declaration argues that those developing AI “should assume their responsibility by working against the risks arising from their technological innovations.” Yet from these various warnings, it is not entirely clear whether it is the people developing AI, or the technology itself, which should be encouraged not to do harm; in other words, whether it is Frankenstein or his monster against whose maleficence we should be guarding. At the heart of this quandary is the question of autonomy.

2.3.3 *Autonomy: The Power to Decide (to Decide)*

When we adopt AI and its smart agency, we willingly cede some of our decision-making power to technological artefacts. Thus, affirming the principle of autonomy in the context of AI means striking a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents. The risk is that the growth in artificial autonomy may undermine the flourishing of human autonomy. It is not therefore surprising that the principle of autonomy is explicitly stated in four of the six documents. The Montreal Declaration articulates the need for a balance between human- and machine-led decision-making, stating that “the development of AI should promote the autonomy [*italics added*] of all human beings”. The EGE argues that autonomous systems “must not impair [the] freedom of human beings to set their own standards and norms,” while AIUK adopts the narrower stance that “the autonomous power to hurt, destroy or deceive human beings should never be vested in AI.” The Asilomar document similarly supports the principle of autonomy, insofar as “humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.” It is therefore clear both that the autonomy of humans should be promoted and that the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected or re-established (consider the case of a pilot able to turn off the automatic pilot and regain full control of the airplane). This introduces a notion we might call ‘meta-autonomy,’ or a ‘decide-to-delegate’ model: humans should retain the power to decide which decisions to take: exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making. Any delegation should also remain overridable in principle (i.e., deciding to decide again).

2.3.4 *Justice: Promoting Prosperity, Preserving Solidarity, Avoiding Unfairness*

The decision to make or delegate decisions does not take place in a vacuum. Nor is this capacity distributed equally across society. The consequences of this disparity in autonomy are addressed in the principle of justice. The importance of ‘justice’ is explicitly cited in the Montreal Declaration, which argues that “the development of AI should promote justice and seek to eliminate all types of discrimination,” while the Asilomar Principles include the need for both “shared benefit” and “shared prosperity” from AI. Under its principle named “Justice, equity and solidarity,” the EGE argues that AI should “contribute to global justice and equal access to the benefits” of AI technologies. It also warns against the risk of bias in datasets used to train AI systems, and—unique among the documents—argues for the need to defend against threats to “solidarity,” including “systems of mutual assistance such as in social insurance and healthcare.” Elsewhere ‘justice’ has still other meanings

(especially in the sense of fairness), variously relating to the use of AI to correct past wrongs such as eliminating unfair discrimination, promoting diversity, and preventing the rise of new threats to justice. The diverse ways in which justice is characterised hints at a broader lack of clarity over AI as a human-made reservoir of ‘smart agency.’ Put simply, are we (humans) the patient, receiving the ‘treatment’ of AI, the doctor prescribing it? Or both? This question can only be resolved with the introduction of a fifth principle which emerges from our analysis.

2.3.5 Explicability: Enabling the Other Principles Through Intelligibility and Accountability

The short answer to the question of whether ‘we’ are the patient or the doctor is that actually we could be either, depending on the circumstances and on who ‘we’ are in everyday life. The situation is inherently unequal: a small fraction of humanity is currently engaged in the development of a set of technologies that are already transforming the everyday lives of almost everyone else. This stark reality is not lost on the authors whose documents we analyze. All of them refer to the need to understand and hold to account the decision-making processes of AI. Different terms express this principle: “transparency” in Asilomar and EGE; both “transparency” and “accountability” in IEEE; “intelligibility” in AIUK; and as “understandable and interpretable” by the Partnership. Each of these principles captures something seemingly novel about AI: that its workings are often invisible or unintelligible to all but (at best) the most expert observers.

The addition of the principle of ‘explicability,’ incorporating both the epistemological sense of ‘intelligibility’ (as an answer to the question ‘how does it work?’) and in the ethical sense of ‘accountability’ (as an answer to the question ‘who is responsible for the way it works?’), is the crucial missing piece of the AI ethics jigsaw. It complements the other four principles: for AI to be beneficent and non-maleficent, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our ‘decision about who should decide’ must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must know whom to hold accountable in the event of a serious, negative outcome, which would require in turn adequate understanding of why this outcome arose.

2.3.6 A Synoptic View

Taken together, these five principles capture every one of the 47 principles contained in the six high-profile, expert-driven documents we analysed. Moreover, each principle is included in almost every statement of principles we analyzed (see

Table 2.1 The five principles in the six documents analysed and their occurrence in three recent documents

	Beneficence	Nonmaleficence	Autonomy	Justice	Explicability
AIUK	•	•	•	•	•
Asilomar	•	•	•	•	•
EGE	•	•	•	•	•
IEEE	•	•			•
Montreal	•	•	•	•	•
Partnership	•	•		•	•
AI4People	•	•	•	•	•
EC HLEG	•	•	•	•	•
OECD	•	•	•	•	•

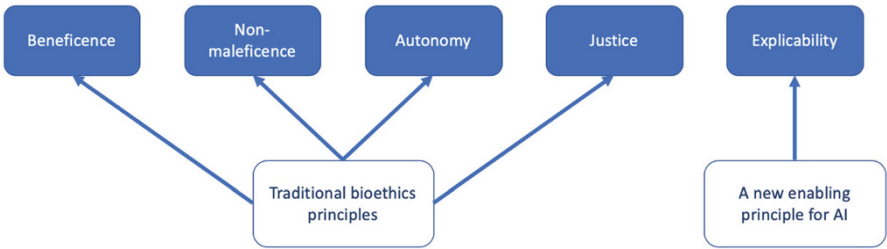


Fig. 2.2 An ethical framework of the five overarching principles for AI which emerged from the analysis

Table 2.1 below). The five principles therefore form an ethical framework within which policies, best practices, and other recommendations may be made. This framework of principles is shown in Fig. 2.2.

2.4 AI Ethics: Whence and for Whom?

It is important to note that each of the six sets of ethical principles for AI that we analyzed emerged either from initiatives with global scope, or from within western liberal democracies. For the framework to be more broadly applicable, it would undoubtedly benefit from the perspectives of regions and cultures presently un- or under-represented in our sample. Of particular interest in this respect is the role of China, which is already home to the world’s most valuable AI start-up (Jezard 2018), enjoys various structural advantages in developing AI (Lee and Triolo 2017), and whose government has stated its ambitions to lead the world in state-of-the-art AI technology by 2030 (China State Council 2017). In its State Council Notice on AI and elsewhere, the Chinese government has expressed interest in further consideration of the social and ethical impact of AI (Ding 2018; Webster et al. 2017). Nor is enthusiasm about the use of technologies unique to governments, but it is also shared

by general publics—more so those in China and India than in Europe or the USA, as new representative survey research shows (Vodafone Institute 2018).

An executive at the major Chinese technology firm Tencent recently suggested that the European Union should focus on developing AI which has “the maximum benefit for human life, even if that technology isn’t competitive to take on [the] American or Chinese market” (Boland 2018). This has been echoed by claims that ethics may be “Europe’s silver bullet” in the “global AI battle” (Delcker 2018). We disagree. Ethics is not the preserve of a single continent or culture. Every company, government agency, and academic institution designing, developing or deploying AI has an obligation to do so in line with an ethical framework along the lines of the one we present here, broadened to incorporate a more geographically, culturally, and socially diverse array of perspectives (Cowls et al. n.d.). Similarly, laws, rules, standards and best practices to constrain or control AI—including all those currently under consideration by regulatory bodies, legislatures and industry groups—would also benefit from close engagement with a unified framework of ethical principles.

2.5 Conclusion: From Principles to Practices

If the framework presented in this article provides a coherent and sufficiently comprehensive overview of the central ethical principles for AI (Floridi et al. 2018), then it can serve as the architecture within which laws, rules, technical standards, and best practices are developed for specific sectors, industries, and jurisdictions. In these contexts, the framework may play both an enabling role (consider, for example, the use of AI to help meet the United Nations Sustainable Development Goals), and a constraining one (as in the need to regulate AI technologies in the context of online crime and cyberwarfare: King et al. 2018; Taddeo and Floridi 2018). Indeed, the framework played a valuable role in the work of AI4People, Europe’s first global forum on the social impact of AI, which recently adopted it to propose 20 concrete recommendations for a ‘Good AI Society’ to the European Commission (Floridi et al. 2018). Since then it has been largely adopted by the Ethics Guidelines for Trustworthy AI published by the European Commission’s High-Level Expert Group on Artificial Intelligence (HLEGAI 2018, 2019), which in turn has influenced the OECD’s Recommendation of the Council on Artificial Intelligence (OECD 2019), reaching 42 countries⁶ (see Table 2.1).

The development and use of AI hold the potential for both positive and negative impact on society, to alleviate or to amplify existing inequalities, to cure old problems, or to cause new ones. Charting the course that is socially preferable will depend not only on well-crafted regulation and common standards, but also on the use of a framework of ethical principles, within which concrete actions can be

⁶<https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm>

situated. We believe that the framework presented here as emerging from the current debate will serve as valuable architecture for securing positive social outcomes from AI technology and move from good principles to good practices (Cowls et al. 2019; Morley et al. 2019).

Disclosure Floridi chaired the AI4People project and Cowls was the rapporteur. Floridi is also a member of the European Commission’s High-Level Expert Group on Artificial Intelligence (HLEGAI).

Funding Floridi’s work was supported by (i) Privacy and Trust Stream—Social lead of the PETRAS Internet of Things research hub—PETRAS is funded by the UK Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1; (ii) Facebook; and (iii) Google. Cowls is the recipient of a Doctoral Studentship from the Alan Turing Institute.

References

- Beauchamp, T.L., and J.F. Childress. 2012. *Principles of biomedical ethics*. Oxford: Oxford University Press.
- Boland, H. 2018. Tencent executive urges Europe to focus on ethical uses of artificial intelligence. *The Telegraph*, October 14. <https://www.telegraph.co.uk/technology/2018/10/14/tencent-executive-urges-europe-focus-ethical-uses-artificial/>
- China State Council. 2017. *State Council notice on the issuance of the next generation Artificial Intelligence development plan*, July 8. Retrieved September 18, 2018, from http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. Translation by Creemers, R., G. Webster, P. Triolo, and E. Kania. <https://www.newamerica.org/documents/1959/translation-fulltext-8.1.17.pdf>
- Corea, F. 2019. *AI knowledge map: How to classify AI technologies, a sketch of a new AI technology landscape*. First appeared in Medium—Artificial Intelligence. https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020. Reproduced in Corea, F. 2019. *An introduction to data*, 26. Springer.
- Cowls, J., L. Floridi, and M. Taddeo. 2018. *The challenges and opportunities of ethical AI*. Artificially Intelligent. https://digitransglasgow.github.io/ArtificiallyIntelligent/contributions/04_Alan_Turing_Institute.html
- Cowls, J., T. C. King, M. Taddeo, and L. Floridi. 2019. *Designing AI for social good: Seven essential factors*. <http://ssrn.com/abstract=3388669>
- Cowls, J., M.-T. Png, and Y. Au. n.d. *Foundations for geographic representation in algorithmic ethics*. Unpublished.
- Delcker, J. 2018. *Europe’s silver bullet in global AI battle: Ethics*. Politico, March 3. <https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/>
- Ding, J. 2018. *Deciphering China’s AI dream*, March. https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf
- European Group on Ethics in Science and New Technologies. 2018. *Statement on Artificial Intelligence, robotics and ‘autonomous’ systems*, March. https://ec.europa.eu/info/news/ethics-artificial-intelligence-statement-ecg-released-2018-apr-24_en
- Floridi, L. 2013. *The ethics of information*. Oxford: Oxford University Press.
- . 2019a. What the near future of Artificial Intelligence could be. *Philosophy & Technology* 32 (1): 1–15. <https://doi.org/10.1007/s13347-019-00345-y>.
- . 2019b. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology* 32 (2): 185–193. <https://doi.org/10.1007/s13347-019-00354-x>.



- Floridi, L., M. Taddeo, and M. Turilli. 2009. Turing's imitation game: Still an impossible challenge for all machines and some judges—An evaluation of the 2008 Loebner contest. *Minds and Machines* 19 (1): 145–150. <https://doi.org/10.1007/s11023-008-9130-6>.
- Floridi, L., J. Cows, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena. 2018. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28 (4): 689–707. <https://doi.org/10.1007/s11023-018-9482-5>.
- Hagendorff, T. 2019. *The ethics of AI ethics—An evaluation of guidelines*. <https://arxiv.org/abs/1903.03425>
- HLEGAI [High Level Expert Group on Artificial Intelligence], European Commission. 2018. *Draft ethics guidelines for trustworthy AI*, December 18. <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>
- . 2019. *Ethics guidelines for trustworthy AI*, April 8. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- House of Lords Artificial Intelligence Committee. 2018. *AI in the UK: Ready, willing and able*, April 16. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm>
- Jezard, A. 2018. *China is now home to the world's most valuable AI start-up*. World Economic Forum, April 11. <https://www.weforum.org/agenda/2018/04/chart-of-the-day-china-now-has-the-worlds-most-valuable-ai-startup/>
- King, T., N. Aggarwal, M. Taddeo, and L. Floridi 2018. *Artificial Intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions*, May 22. <https://ssrn.com/abstract=3183238>
- Lee, K., and P. Triolo 2017. *China's Artificial Intelligence revolution: Understanding Beijing's structural advantages*. Eurasian Group, December. <https://www.eurasiagroup.net/live-post/ai-in-china-cutting-through-the-hype>
- McCarthy, J., M.L. Minsky, N. Rochester, and C.E. Shannon. 2006. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine* 27 (4): 12. <https://doi.org/10.1609/aimag.v27i4.1904>.
- Montreal Declaration for a Responsible Development of Artificial Intelligence. 2017. *Announced at the conclusion of the Forum on the Socially Responsible Development of AI*, November 3. <https://www.montrealdeclaration-responsibleai.com/the-declaration>
- Morley, J., L. Floridi, L. Kinsey, and A. Elhalal. 2019. *From what to how. An overview of AI ethics tools, methods and research to translate principles into practices*. ArXiv:1905.06876 [Cs]. Retrieved from <http://arxiv.org/abs/1905.06876>
- OECD. 2019. *Recommendation of the Council on Artificial Intelligence*. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- Partnership on AI. 2018. *Tenets*. <https://www.partnershiponai.org/tenets/>
- Samuel, A.L. 1960. Some moral and technical consequences of automation—A refutation. *Science* 132 (3429): 741–742. <https://doi.org/10.1126/science.132.3429.741>.
- Taddeo, M., and L. Floridi. 2018. Regulate artificial intelligence to avert cyber arms race. *Nature* 556 (7701): 296–298.
- The IEEE Initiative on Ethics of Autonomous and Intelligent Systems. 2017. *Ethically aligned design*, v2. <https://ethicsinaction.ieee.org>
- Turing, A.M. 1950. Computing machinery and intelligence. *Mind* 5 (236): 433–460. <https://doi.org/10.1093/mind/lix.236.433>.
- Vodafone Institute for Society and Communications. 2018. *New technologies: India and China see enormous potential—Europeans more sceptical*. <https://www.vodafone-institut.de/digitising-europe/digitisation-india-and-china-see-enormous-potential/>
- Webster, G., R. Creemers, P. Triolo, and E. Kania. 2017. *China's plan to 'lead' in AI: Purpose, prospects, and problems*. New America, August, 1. <https://www.newamerica.org/cybersecurity-initiative/blog/chinas-plan-lead-ai-purpose-prospects-and-problems/>
- Wiener, N. 1960. Some moral and technical consequences of automation. *Science* 131 (3410): 1355–1358. <https://doi.org/10.1126/science.131.3410.1355>.

Yang, G.Z., J. Bellingham, P.E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, B.J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. Veloso, Z.L. Wang, and R. Wood. 2018. The grand challenges of science robotics. *Science robotics* 3 (14): eaar7650. <https://doi.org/10.1126/scirobotics.aar7650>.

Chapter 3

An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations



Luciano Floridi , Josh Cows , Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena

Abstract This article reports the findings of AI4People, a year-long initiative designed to lay the foundations for a “Good AI Society”. We introduce the core opportunities and risks of AI for society; present a synthesis of five ethical principles that should undergird its development and adoption; and offer 20 concrete

L. Floridi (✉)

Oxford Internet Institute, University of Oxford, Oxford, UK
e-mail: luciano.floridi@oii.ox.ac.uk

J. Cows

Oxford Internet Institute, University of Oxford, Oxford, UK

Alan Turing Institute, London, UK

e-mail: Josh.cows@oii.ox.ac.uk

M. Beltrametti

Naver Corporation, Grenoble, France

e-mail: Monica.beltrametti@naverlabs.com

R. Chatila

French National Center of Scientific Research, Paris, France

Institute of Intelligent Systems and Robotics at Pierre, Marie Curie University, Paris, France

e-mail: Raja.chatila@sorbonne-universite.fr; chatila@isir.upmc.fr

P. Chazerand

Digital Europe, Brussels, Belgium

e-mail: patrice.chazerand@digitaleurope.org

V. Dignum

Department of Computing Science, University of Umeå, Umeå, Sweden

Delft Design for Values Institute, Delft University of Technology, Delft, the Netherlands

e-mail: virginia@cs.umu.se

C. Luetge

TUM School of Governance, Technical University of Munich, Munich, Germany

e-mail: luetge@tum.de

recommendations – to assess, to develop, to incentivise, and to support good AI – which in some cases may be undertaken directly by national or supranational policy makers, while in others may be led by other stakeholders. If adopted, these recommendations would serve as a firm foundation for the establishment of a Good AI Society.

Keywords Artificial intelligence · AI4People · Data Governance · Digital Ethics · Governance · Ethics of AI

3.1 Introduction

AI is not another utility that needs to be regulated once it is mature. It is a powerful force, a new form of smart agency, which is already reshaping our lives, our interactions, and our environments. AI4People was set up to help steer this powerful force towards the good of society, everyone in it, and the environments we share. This White Paper is the outcome of the collaborative effort by the AI4People Scientific Committee – comprising 12 experts and chaired by Luciano Floridi¹ – to propose a series of recommendations for the development of a Good AI Society.

¹Besides Luciano Floridi, the members of the Scientific Committee are: Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Josh Cowsls is the rapporteur. Thomas Burri contributed to an earlier draft.

R. Madelin

Defence Science and Technology Laboratories, Salisbury, UK

Centre for Technology and Global Affairs, University of Oxford, Oxford, UK

e-mail: Robert.Madelin@ec.europa.eu; robert.madelin@fipa.com

U. Pagallo

Department of Law, University of Turin, Turin, Italy

e-mail: ugo.pagallo@unito.it

F. Rossi

IBM Research, Albany, NY, USA

University of Padova, Padova, Italy

e-mail: Francesca.Rossi2@ibm.com

B. Schafer

School of Law, University of Edinburgh Law School, Edinburgh, UK

e-mail: B.Schafer@ed.ac.uk

P. Valcke

Centre for IT & IP Law, Catholic University of Leuven, Leuven, Flanders, Belgium

Bocconi University, Milan, Italy

e-mail: peggy.valcke@kuleuven.be

E. Vayena

Bioethics, Health Ethics and Policy Lab, ETH Zurich, Zurich, Switzerland

e-mail: effy.vayena@hest.ethz.ch