PhD course: Ethics of Artificial Intelligence
University of L'Aquila

First Part
Donatella Donati – 12 hours

Objectives

The aim of the module is to explore some the main ethical issues in the design and use of both AI systems and information and communication technologies (henceforth, ICTs).

Understanding the role of moral values in AI and ICTs is indispensable to their design and their use. Questions on the moral status of artificial agents, the nature of human-artificial agents relationship, and production, access, and control of information will be at the heart of moral challenges surrounding both AI systems and ICTs.

We will examine ethical theories and practices from philosophical and interdisciplinary perspectives relating to the design and use of AI systems and ITCs.

Contents

The module will address the following topics:

Moral Machines
Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-Fançois Bonnefon, and Iyad Rahwan, 'The Moral Machine Experiment', Nature 563 (2018): 59-64.
Amitai Etzioni and Oren Etzioni, 'Incorporating Ethics into Artificial Intelligence', Journal of Ethics 21 (2017): 403-18.
Peter Railton, 'Ethical Learning, Natural and Artificial', in S. Matthew Liao (ed), Ethics of Artificial Intelligence (OUP, 2020).

Moral Status and Rights of Artificial Agents
John Basl and Joseph Bowen, 'AI as a Moral Right-Holder', in M. Dubber, F. Pasquale, and S. Das (eds), The Oxford Handbook of Ethics of AI (OUP, 2019)

S. Matthew Liao, 'The Moral Status and Rights of Artificial Intelligence', in S. Matthew Liao (ed), Ethics of Artificial Intelligence (OUP, 2020)
Eric Schwitzgebel and Mara Garza, 'Designing AI with Rights, Counsciousness, Self-Respect, and Freedom', in S. Matthew Liao (ed), Ethics of Artificial Intelligence (OUP, 2020).

Autonomy and Manipulation
Richard J. Arneson, 'Nudge and Shove', Social Theory and Practice, 41 (2015): 668-691.
Sarah Buss, 'Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints', Ethics, 115 (2005): 195–235.
Robert Noggle, 'Manipulative Actions: A Conceptual and Moral Analysis', American Philosophical Quarterly, 33 (1996): 43-55.

Algorithmic bias and discrimination
Joy Buolamwini, J. and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparity in Commercial Gender Classification', Proceedings of Machine Learning Research 81 (2018): 1-15. <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> [Accessed on 13 June 2019].
Timnit Gebru, 'Race and Gender', in M. Dubber, F. Pasquale, and S. Das (eds), The Oxford Handbook of Ethics of AI (OUP, 2020)
Cass Sunstein, 'Algorithms, Correcting Biases.' Social Research 86 (2019): 499-511.

Privacy and Surveillance
Andrei Marmor, 'What is the Right to Privacy?', Philosophy and Public Affairs 43 (2015): 3-26.
James Rachels, 'Why Privacy is Important', Philosophy and Public Affairs, 4 (1975): 323–33.
Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting (Oxford UP, 2016), ch.8 ("Surveillance and the Examined Life: Cultivating the Technomoral Self in a Panoptic World")

Robot-Human Relationship
Judith Donath, 'Ethical Issues in Our Relationship with Artificial Entities', in M. Dubber, F. Pasquale, and S. Das (eds), The Oxford Handbook of Ethics of AI (OUP, 2020).
Paul Formosa, 'Robot Autonomy vs. Human Autonomy: Social Robots, Artificial Intelligence (AI), and the Nature of Autonomy', Minds and Machines (2021) 31: 595-616.
Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting (Oxford UP, 2016), ch.9 ("Robots at War and at Home: Preserving the Technomoral Virtues of Care and Courage").

Teaching Methods

The course consists of six two-hour seminars.

The best way to learn philosophy is by doing it, so the students will have to actively engage in discussion, by presenting ideas, questions, objections. You cannot just sit back and talk about past debates – ethics of AI and ICTs is a fast-moving fields: the debates are happening now and students are expected to participate in them.

Before the meetings, students are expected to: (i) have done the essential reading and (ii) have thought of at least one question about the essential reading to bring to class. During the seminar we will work together to understand the reading and assess the claims and arguments.

The essential readings will be available on Microsoft Teams.

****

Second Part by Abeer Dyoub, Univaq – 6 hours

Objectives: This Part will concentrate on approaches, aspects and issues related to computing machine ethics.

Contents:

Introduction of different existing approaches for implementing ethics into intelligent Artificial Agents.
Discussion of the relation between machine ethics and explainability, and the role of logic programming.
Discussion of some of the challenges and limitations of implementing machine ethics.
Finally, introduction to our proposal for machine ethics (what we did/our contribution).

References:

James H. Moor. "The Nature, Importance, and Difficulty of Machine Ethics". In: IEEE Intelligent Systems 21.4 (2006), pp. 18–21. doi:10.1109/MIS.2006.80.

Tom L Beauchamp and James F Childless. "Principles of Biomedical Ethics". In: International Clinical Psychopharmacology 6.2 (1991), pp. 129–130. doi: 10.1001/jama.1984.03340360075041.

Stefania Costantini et al. “Trustworthiness and Safety for Intelligent Ethical Logical Agents via Interval Temporal Logic and Runtime Self-Checking”. In: 2018 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 26-28, 2018. AAAI Press, 2018.

Wendell Wallach and Colin Allen. Moral machines: Teaching robots right from wrong. Oxford University Press, 2008.

Wendell Wallach, Colin Allen, and Iva Smit. “Machine morality: bottom- up and top-down approaches for modelling human moral faculties”. In: AI Soc. 22.4 (2008), pp. 565–582. doi: 10.1007/s00146-007-0099-0.

Miles Brundage. “Limitations and risks of machine ethics”. In: J. Exp. Theor. Artif. Intell. 26.3 (2014), pp. 355–372. doi: 10.1080/0952813X.2014.895108. https://doi.org/10.1080/0952813X.2014.895108.

Susan Leigh Anderson. “Asimov’s "three laws of robotics" and machine metaethics”. In: AI Soc. 22.4 (2008), pp. 477–493. doi: 10 .1007/s00146-007-0094-5.

Daniel M Bartels. “Principled moral sentiment and the flexibility of moral judgment and decision making”. In: Cognition 108.2 (2008), pp. 381–417.

Luís Moniz Pereira, The Anh Han, António Barata Lopes: Employing AI to Better Understand Our Morals. Entropy 24(1): 10 (2022)

Luís Moniz Pereira: The carousel of ethical machinery. AI Soc. 36(1): 185-196 (2021)

Luís Moniz Pereira: Should I kill or rather not? AI Soc. 34(4): 939-943 (2019)

Further references will be provided during the course.