Seminar: Ethics of AI // Philosophische und ethische Aspekte künstlicher Intelligenz

Summer Term 2021

JProf. Dr. Felix Lindner (Juniorprofessor for Explainable Artificial Intelligence, Ulm University)
Prof. Dr. Bernhard Nebel (Foundations of Artificial Intelligence, University of Freiburg)
Prof. Dr. Oliver Müller (Philosophy of Technology, University of Freiburg)

A significant subset of the articles are only available from within the university network. To download, you must use a university computer or VPN.

Topics

Area A*: Meta-Ethics*

A1: *Machines as artificial moral agents*

- Misselhorn, C.: Artificial Morality. Concepts, Issues and Challenges. Society 55 (2018), 161-169.

A2: *Machines as artificial moral patients*

- Gunkel, D.J.: The machine question. Critical perspectives in AI, robots, and ethics. Cambridge MA 2017, 93-157. (PDF)

A3: *Should robots have rights?*

- Coeckelbergh, M.: Robot rights? Towards a social-relational justification of moral considerations. Ethics and Information Technology 12:209--221 (2010) (PDF)

A4: *Should robots have no rights?*

- Bryson, J.J.: Robots Should be Slaves. Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues, Chapter 11. John Benjamins Publishing Company, 2010, pp. 63--74. (PDF)

Area *B: Machine Ethics*

B1: *What is machine ethics?*

- Bendel, O.: Die Maschinenethik als neues interdisziplinäres Forschungsfeld. In: Liggieri, K. & Müller, O. (eds): Mensch-Maschine-Interaktion. Metzler-Handbuch. Geschichte - Kultur - Ethik. Stuttgart 2019, 361-367.

B2: *What is robot ethics?*

- Loh, J.: Arbeitsfelder der Roboterethik. In: Liggieri, K. & Müller, O. (eds): Mensch-Maschine-Interaktion. Metzler-Handbuch. Geschichte - Kultur - Ethik. Stuttgart 2019, 352-360.

B3: *Utilitarian and Kantian machines*

- Powers, T.M.: Prospects for a Kantian Machine. In: Anderson, M., Anderson, S.L. (eds.): Machine ethics. Cambridge 2011, 464--475. (PDF)
- Cloos, C.: The Utilibot Project. An autonomous mobile robot based on utilitarianism. American Association for Artificial Intelligence 2005. (PDF)

B4*: Formal Ethics*

- Martin Mose Bentzen (2016) The Principle of Double Effect Applied to Ethical Dilemmas of Social Robots. Proceedings of Robophilosophy 2016/TRANSOR 2016. pp 268–279. http://www.hera-project.com/wp-content/uploads/2018/08/bentzen-robophil2016-.pdf

B5: *Empathic and Emotional machines*

- Misselhorn, C.: Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co. Stuttgart 2021.

B6*: Superintelligence as a danger to humankind*

- Nick Bostrom (2012) The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds & Machines 22:71–85 https://link.springer.com/content/pdf/10.1007/s11023-012-9281-3.pdf

Area *C: Core Concepts in AI Ethics*

C1: *Responsibility*

- Gunkel, D. J.: Mind the gap: responsible robotics and the problem of responsibility. Ethics and Information Technology 1--14 (2017) (PDF)

C2: *Trust*

- Coeckelbergh, M.: Can we trust robots? Ethics and Information Technology 14 (2012), 53--60. (PDF)
- Arjen van Witteloostuijn: A Game-Theoretic Framework of Trust, International Studies of Management & Organization, Vol. 33, No. 3, 53-71, 2003. (PDF)

C3: *Trust from a game-theoretic point of view*

- Erin Paeng, Jane Wu, James C. Boerkoel: Human-Robot Trust and Cooperation Through a Game Theoretic Framework. AAAI 2016: 4246-4247. (PDF)

C4: *Fairness*

- Leben, D.: Normative Principles for Evaluating Fairness in Machine Learning. Proceedings of the 2020 Conference on AI, Ethics, and Society (2020). https://www.aies-conference.com/2020/wp-content/papers/051.pdf

Area *D: Applied AI Ethics*

D1: *Guidelines for AI research*

- Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Stephen Cave: The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. (PDF)
- European Commission: Ethics guidelines for trustworthy AI

D2: *Biases in machine learning*

- Benthall, S., Haynes, B. D.: Racial categories in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 289–298 (2019) (PDF)

D3: *Autonomous weapons systems*

- Lim, D.: Killer robots and human dignity. In: The Second AAAI / ACM Annual Conference on AI, Ethics, and Society (2019) (PDF)
- Altmann, J. (2019) Autonomous Weapon Systems Dangers and Need for an International Prohibition, in C. Benzmüller, H. Stuckenschmidt (eds.), KI 2019: Advances in Artificial Intelligence 42nd German Conference on AI, Kassel, Germany, September 2326, 2019, Proceedings, Cham: Springer, 2019, 1-17.

D4: *Autonomous cars*

- https://www.moralmachine.net/
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature 563:59--65 (2018) (PDF)