# Ethics of AI Explained

**Clayton Peterson**
Université du Québec à Trois-Rivières
3351 Bd des Forges, Trois-Rivières (QC), G8Z 4M3
clayton.peterson@uqtr.ca

### Abstract

This tutorial introduces participants to the main issues and themes pertaining to ethics of artificial intelligence (AI). Analyzing what ethics of AI is by reflecting on our understanding of both ethics and AI, the aim is to clarify and expose how ethics of AI can be conceived in different ways depending on the approach one adopts. Through the use of concrete examples, participants will be introduced to various issues in the ethics of AI literature, allowing them to reflect upon their own research and practice.

## Bio

Clayton Peterson is an associate professor of philosophy at the Department of philosophy and arts at Université du Québec à Trois-Rivières. Chairholder of the UQTR Research chair in ethics of AI and vice-president of the research ethics committee – psychology and psychoeducation of his institution, he holds a Ph.D. in philosophy from Université de Montréal and did his postdoctoral studies at the Munich Center for Mathematical Philosophy (Ludwig-Maximilians Universität München). His work concentrates on the theoretical and practical limitations of automating ethical behavior and reasoning as well as on the integration of ethics to computer science and engineering.

## Introduction

Along with the technological growth we are experiencing comes deep ethical concerns regarding not only regulation and usage of AI systems, but also research and development. As new technologies and techniques become available, people are realizing the necessity of reflecting on the ethics of AI as well as on the principles that should guide our endeavors and the consequences we should aim to avoid.

## Target Audience and Objective

The aim of this tutorial is to promote intersectoral reflections on AI and introduce participants to the ethics of AI. Starting from the assumption that understanding the ethics of AI requires a proper understanding of both ethics and AI, we reflect on what both ethics and AI are in order to bring to light different ways in which ethics of AI can be understood, emphasizing the limitations of ethical AI and its relationship to explainable AI (XAI). This tutorial targets a multidisciplinary audience broadly composed of scholars from exact sciences as well as from the humanities, with a specific interest for scholars from computer science, engineering, and philosophy.

## Relevance

This tutorial builds on the idea that ethics of AI needs to be grounded on a dialog between natural science, engineering as well as human and social sciences in order to really contribute to the development of AI. As such, it starts from the assumption that human and social sciences should not only inform technological developments but should also be informed by natural sciences and engineering. Scholars in human and social sciences need to understand the constraints and limitations surrounding technological developments to be able to integrate these concerns into their ethical and social reflections. They need to understand how the values they promote can be integrated within technological developments in order to form an informed judgment. Otherwise, human and social sciences will have a very limited impact on technological developments. Ethical concerns, if they are not informed by the practice, are essentially theoretical and void of any concrete empirical consequence. As such, although theoretical considerations regarding the principles and values that should guide technological developments are important, there is a pressing need to stop reflecting upon the theory and think of feasible ways to apply this theory to concrete cases. This can be understood as a shift of perspective from normative ethics to applied ethics. By emphasizing conceptual distinctions regarding both ethics and AI, this tutorial will expose how ethics of AI can be understood in distinct and complementary ways, from a normative understanding of ethics of AI to an applied one.

## Structure and Scope

1. What is ethics?
   (a) Ethics, truth, and relativism
   (b) Ethical pluralism
   (c) Normative and applied ethics

(d) Are technologies and algorithms value neutral?

2. What is AI?
   (a) AI singularity and AGI
   (b) Applied AI
3. What is ethics of AI?
   (a) Normative ethics and AI singularity
   (b) Normative ethics and applied AI
   (c) Applied ethics and applied AI
4. Ethics, AI, Automation, and XAI
   (a) Pluralism and the possibility of automating ethics to solve dilemmas
   (b) Explainability of automated ethical reasoning and behavior
   (c) Ethical XAI

## Acknowledgments

This work was financially supported by the UQTR Research Chair in the Ethics of AI as well as by the *Fonds de Recherche du Québec* [2023-NP-310505]. It was further supported in part by funding from the Social Sciences and Humanities Research Council.

## References

Adadi, A., and Berrada, M. 2018. Peeking inside the blackbox: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6:52138–52160.

Anderson, M., and Anderson, S. L., eds. 2011. *Machine Ethics*. Cambridge University Press.

Batarseh, F. A.; Freeman, L.; and Huang, C.-H. 2021. A survey on artificial intelligence assurance. *Journal of Big Data* 8(1):1–30.

Biran, O., and Cotton, C. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8. 8–13.

Brundage, M. 2014. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence* 26(3):355–372.

Campbell, J. K.; O'Rourke, M.; and Shier, D., eds. 2004. *Freedom and Determinism*. A Bradford Book.

De Cremer, D., and Kasparov, G. 2022. The ethical AI paradox: Why better technology needs more and not less human responsibility. *AI and Ethics* 2(1):1–4.

Dignum, V. 2019. *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

Dignum, V. 2021. The role and challenges of education for responsible AI. *London Review of Education* 19(1):1–11.

Dubber, M. D.; Pasquale, F.; and Das, S., eds. 2020. *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Eden, A.; Moor, J.; Søraker, J.; and Steinhart, E., eds. 2012. *Singularity Hypotheses*. The Frontiers Collection. Springer.

Etienne, H. 2022. When AI ethics goes astray: A case study of autonomous vehicles. *Social Science Computer Review* 40(1):236–246.

Faden, R. R., and Beauchamp, T. L. 1986. *A history and theory of informed consent*. Oxford University Press.

Floridi, L., ed. 2010. *The Cambridge handbook of information and computer ethics*. Cambridge University Press.

Harman, G. 1978. What is moral relativism? In Goldman, A. I., and Kim, J., eds., *Values and morals*, volume 13 of *Philosophical Studies Series in Philosophy*. Springer. 143–161.

Johnson, D. G. 2009. *Computer ethics*. Prentice Hall, 4th edition.

Kaplan, A., and Haenlein, M. 2019. Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implica tions of artificial intelligence. *Business Horizons* 62(1):15–25.

LaCroix, T. 2022. Moral dilemmas for moral machines. *AI and Ethics* 2(4):737–746.

McCarthy, J., and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Metzer, B., and Michie, D., eds., *Machine Intelligence 4*. Edinburgh University Press. 463–502.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Miller, B. 2021. Is technology value-neutral? *Science, Technology, & Human Values* 46:53–80.

Moor, J. H. 1999. Just consequentialism and computing. *Ethics and Information Technology* 1(1):61–65.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Moore, G. E. 1959. *Principia Ethica [1903]*. Cambridge University Press.

Müller, V. C., ed. 2013. *Philosophy and theory of artificial intelligence*. Springer.

Noble, S. U. 2018. *Algorithms of oppression*. New York University Press.

Peterson, C., and Broersen, J. 2024. Understanding the limits of explainable ethical AI. *International Journal on Artificial Intelligence Tools*.

Peterson, C., and Hamrouni, N. 2022. Preliminary thoughts on defining $f(x)$ for ethical machines. *The International FLAIRS Conference Proceedings* 35.

Peterson, C. 2023. Further thoughts on defining $f(x)$ for ethical machines: Ethics, rational choice, and risk analysis. *The International FLAIRS Conference Proceedings* 36.

Russell, S., and Norvig, P. 2022. *Artificial Intelligence: A Modern Approach*. Global Edition, 4th edition.

Ryan, M. 2020. In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 26:2749–2767.

Satisky, J. 2019. A Duke study recorded thousands of students' faces. Now they are being used all over the world. *The Duke Chronicle*.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3):417–424.

Sen, A., and Williams, B. 1982. *Utilitarianism and Beyond*. Cambridge University Press.

Tolmeijer, S.; Kneer, M.; Sarasua, C.; Christen, M.; and Bernstein, A. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)* 53(6):1–38.

UNESCO. 2021. *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

Van de Poel, I., and Royakkers, L. 2011. *Ethics, technology, and engineering: An introduction*. John Wiley & Sons.

von Braun, J.; Archer, M. S.; Reichberg, G. M.; and Sorondo, M. S., eds. 2021. *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Springer.

Wallach, W., and Allen, C. 2009. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Weinstock, D. 2017. Compromise, pluralism, and deliberation. *Critical Review of International Social and Political Philosophy* 20(5):636–655.

Wooldridge, M. 2021. *A brief history of artificial intelligence: What it is, where we are, and where we are going*. Flatiron Books.