REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *CoRR* abs/1606.06565 (2016).

[2] Guido Boella, Leendert W.N. van der Torre, and Harko Verhagen. 2007. Introduction to Normative Multiagent Systems. In *Normative Multi-agent Systems*.

[3] Nick Bostrom and Eliezer Yudkowsky. 2011. Ethics of Artificial Intelligence. *Cambridge Handbook of Artificial Intelligence* (2011).

[4] Ryan Calo. 2017. Artificial Intelligence Policy: A Primer and Roadmap. https://doi.org/10.2139/ssrn.3015350

[5] David Cooper. 1993. *Value pluralism and ethical choice.* St. Martin Press, Inc.

[6] R M Dawes. 1980. Social Dilemmas. *Annual Review of Psychology* 31, 1 (1980), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125 arXiv:https://doi.org/10.1146/annurev.ps.31.020180.001125

[7] Benedictus de Spinoza. 1883. *A Theologico-Political Treatise.* Dover Publications.

[8] Frank Dignum. 1999. Autonomous Agents with Norms. *Artif. Intell. Law* 7, 1 (1999), 69–79.

[9] F. Dignum. 1999. Autonomous Agents with Norms. *Artificial Intelligence and Law, 7: 69* (1999). https://doi.org/10.1023/A:1008315530323

[10] A. M. Fink. 1964. Equilibrium in a stochastic n-person game. *J. Sci. Hiroshima Univ. Ser. A-I Math.* 28, 1 (1964), 89–93. https://doi.org/10.32917/hmj/1206139508

[11] William K. Frankena. 1973. *Ethics, 2nd edition.* Englewood Cliffs, N.J. : Prentice-Hall,.

[12] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Venable, and Brian Williams. 2016. Embedding Ethical Principles in Collective Decision Support Systems. (2016).

[13] Sven Ove Hansson. 2001. *The structure of values and norms.* Cambridge University Press.

[14] Garrett Hardin. 1968. The Tragedy of the Commons. *Science* 162, 3859 (1968), 1243–1248. https://doi.org/10.1126/science.162.3859.1243 arXiv:https://science.sciencemag.org/content/162/3859/1243.full.pdf

[15] Thomas Hobbes. 1651. *Leviathan, 1651.* Menston, Scolar P.

[16] Robert L. Holmes. 1990. The Limited Relevance of Analytical Ethics to the Problems of Bioethics. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 15, 2 (04 1990), 143–159. https://doi.org/10.1093/jmp/15.2.143 arXiv:http://oup.prod.sis.lan/jmp/article-pdf/15/2/143/2681996/15-2-143.pdf

[17] Terry Horgan and Mark Timmons. 2010. Untying a knot from the inside out: Reflections on the "paradox" of supererogation. *Social Philosophy and Policy* 27 (07 2010), 29 – 63. https://doi.org/10.1017/S026505250999015X

[18] Junling Hu and Michael P. Wellman. 2003. Nash Q-learning for General-sum Stochastic Games. *J. Mach. Learn. Res.* 4 (Dec. 2003), 1039–1069. http://dl.acm.org/citation.cfm?id=945365.964288

[19] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *NeurIPS*.

[20] Lawrence Kohlberg, Charles Levine, and A. Hewer. 1983. Moral Stages: a Current Formulation and a Response to Critics.

[21] Peter Kollock. 1998. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183 arXiv:https://doi.org/10.1146/annurev.soc.24.1.183

[22] B. De Schutter L. Busoniu, R. Babuska. 2010. Multi-agent reinforcement learning: An overview. *Innovations in Multi-Agent Systems and Applications – 1* (2010), 183–221.

[23] Joel Z. Leibo, Vinícius Flores Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. *CoRR* abs/1702.03037 (2017). arXiv:1702.03037 http://arxiv.org/abs/1702.03037

[24] Michael L. Littman. 1994. Markov Games As a Framework for Multi-agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning (ICML'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 157–163. http://dl.acm.org/citation.cfm?id=3091574.3091594

[25] John Locke. 1967. *Two Treatises of Government.* Cambridge: Cambridge University Press.

[26] Javier Morales, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, Wamberto Vasconcelos, and Michael Wooldridge. 2015. Online automated synthesis of compact normative systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 10, 1 (2015), 33.

[27] Javier Morales, Maite López-Sánchez, Juan Antonio Rodríguez-Aguilar, Michael Wooldridge, and Wamberto W. Vasconcelos. 2015. Synthesising Liberal Normative Systems. *Proceedings of the fourteenth International Conference on Autonomous Agents and Multiagent Systems, Wiley* (2015).

[28] Gonçalo Neto. 2005. From Single-Agent to Multi-Agent Reinforcement Learning: Foundational Concepts and Methods. (2005).

[29] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. *Version 2. IEEE* (2017).

[30] John Rawls. 1958. Justice as Fairness. *Philosophical Review* 67, 2 (1958), 164–194. https://doi.org/10.2307/2182612

[31] Jean-Jacques Rousseau. 1950. *The Social Contract.* New York: Harmondsworth, Penguin.

[32] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems* 7 (2011), 21–54.

[33] Marc Serramia, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, Javier Morales, Michael Wooldridge, and Carlos Ansotegui. 2018. Exploiting moral values to choose the right norms. In *Proceedings of the 1st Conference on artificial intelligence, ethics and society (AIES'18)*. 1–7. https://doi.org/10.1145/3278721.3278735

[34] Marc Serramia, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, Manel Rodriguez, Michael Wooldridge, Javier Morales, and Carlos Ansotegui. 2018. Moral Values in Norm Decision Making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18)*. International Foundation for Autonomous Agents and Multiagent Systems, 1294–1302.

[35] Yoav Shoham and Kevin Leyton-Brown. 2009. *Multiagent Systems - Algorithmic, Game-Theoretic, and Logical Foundations.* Cambridge University Press.

[36] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction.* MIT Press. http://www.worldcat.org/oclc/37293240

[37] James O. Urmson. 1958. Saints and Heroes. In *Essays in Moral Philosophy*, A. I. Melden (Ed.). University of Washington Press.

[38] Ibo van de Poel and Lambèr Royakkers. 2011. *Ethics, Technology, and Engineering: An Introduction.* Wiley-Blackwell.

[39] Wendell Wallach. 2008. Implementing Moral Decision Making Faculties in Computers and Robots. *AI and Society* 22, 4 (2008), 463–475. https://doi.org/10.1007/s00146-007-0093-6

[40] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez Guzmán, and Joel Z. Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 683–692. http://dl.acm.org/citation.cfm?id=3306127.3331756

[41] Christopher J. C. H. Watkins and Peter Dayan. 1992. Technical Note Q-Learning. *Machine Learning* 8 (1992), 279–292. https://doi.org/10.1007/BF00992698

[42] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building Ethics into Artificial Intelligence. In *IJCAI*.