Literature sources

[The methodological literature particularly central to the seminar is highlighted in bold.]

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Propublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Asimov, I. (2004). *I, Robot*. Random House Worlds.

Barocas, S., Hardt, M., & Narayanan, A. (2021). *Fairness and Machine Learning*. 253.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Bartneck C, Lütge C, Wagner A, et al. (2021) *An Introduction to Ethics in Robotics and AI*. SpringerBriefs in Ethics. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-51110-4

Behrendt, H., & Loh, W. (2022). Informed consent and algorithmic discrimination – is giving away your data the new vulnerable? *Review of Social Economy*, *80*(1), 58–84. https://doi.org/10.1080/00346764.2022.2027506

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bolukbasi et al. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Proceedings of the International Conference on Advances in Neural Information Processing Systems.

Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*.

Brandner, L. T., & Hirsbrunner, S. D. (2023). Algorithmische Fairness in der polizeilichen Ermittlungsarbeit:Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung. *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, *32*(1), Article 1. https://doi.org/10.14512/tatup.32.1.24

Carroll, J. M. (1999). Five Reasons for Scenario-Based Design. *Nd Hawaii International Conference on System Sciences*, 12.

Cheong, J., Kalkan, S., & Gunes, H. (2023). Counterfactual Fairness for Facial Expression Recognition. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Computer Vision – ECCV 2022 Workshops* (Vol. 13805, pp. 245–261). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25072-9_16

Collingridge, D. (1982). *The Social Control of Technology*. Eweb:40054. https://repository.library.georgetown.edu/handle/10822/792071

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*.

Crawford K and Calo R (2016) There is a blind spot in AI research. *Nature* 538(7625): 311–313. DOI: 10.1038/538311a.

Crawford, K., & Joler, V. (2019). Anatomy of an AI System. *Virtual Creativity*, *9*(1), 117–120. https://doi.org/10.1386/vcr_00008_7

European Commission (2021) Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM(2021) 206 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed 9 August 2021).

Fischer, M. T., Hirsbrunner, S. D., Jentner, W., Miller, M., Keim, D. A., & Helm, P. (2022). Promoting Ethical Awareness in Communication Analysis: Investigating Potentials and Limits of Visual Analytics for Intelligence Applications. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 877–889. https://doi.org/10.1145/3531146.3533151

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). *CapAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act* (SSRN Scholarly Paper No. 4064091). https://doi.org/10.2139/ssrn.4064091

Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16–23. https://doi.org/10.1145/242485.242493

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Friedman, B., & Kahn Jr, P. H. (2007). Human values, ethics, and design. In *The human-computer interaction handbook* (pp. 1267–1292). CRC press.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Henning, T. (2019). *Allgemeine Ethik*. UTB.

Hirsbrunner, S. D., Tebbe, M., & Müller-Birn, C. (2022). From critical technical practice to reflexive data science. *Convergence: The International Journal of Research into New Media Technologies*, *Special Issue: Critical Technical Practice(s) in Digital Research*. https://doi.org/10.1177/13548565221132243

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Lum, K., & Isaac, W. (2016). To Predict and Serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Lütge, C., Hohma, E., Boch, A., Poszler, F., & Corrigan, C. (2022). *On a Risk-Based Assessment Approach to AI Ethics Governance*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mühlhoff, R. (2023). Predictive privacy: Collective data protection in the context of artificial intelligence and big data. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231166886

Nathan, L. P., Klasnja, P. V., & Friedman, B. (2007). Value scenarios: A technique for envisioning systemic effects of new technologies. *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, 2585–2590. https://doi.org/10.1145/1240866.1241046

Rössler, B. (2001). *Der Wert des Privaten*. Suhrkamp Frankfurt am Main.

Taylor, L., & Floridi, L. (2017). Group Privacy: New Challenges of Data Technologies. *Group Privacy*.

Rosson, M. B., & Carroll, J. M. (2007). Scenario-based design. In *The Human-Computer Interaction Handbook* (pp. 1067–1086). CRC Press.

Spindler, M., Booz, S., Gieseler, H., Runschke, S., Wydra, S., & Zinsmaier, J. (2020). How to achieve integration? In B. Gransche & A. Manzeschke (Eds.), *Das geteilte Ganze: Horizonte Integrierter Forschung für künftige Mensch-Technik-Verhältnisse* (pp. 213–239). Springer Fachmedien. https://doi.org/10.1007/978-3-658-26342-3_11

Stanford Encyclopedia of Philosophy 2019: https://plato.stanford.edu/entries/value-intrinsic-extrinsic/#WhaHasIntVal

Thomas WI and Thomas DST (1928) *The Child in America; Behavior Problems and Programs*. New York: A. A. Knopf. Available at: https://catalog.hathitrust.org/Record/003925082 (accessed 29 September 2019).

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Werner, M. H. (2021). *Einführung in die Ethik*. J.B. Metzler. https://doi.org/10.1007/978-3-476-05293-3