Alignment

Readings on the moral alignment between human values and current data science practices.

Title	Citation
The Alignment Problem: Machine Learning and Human Values	Christian (2021)
Artificial Intelligence, Values, and Alignment	Gabriel (2020)
Living Well Together Online: Digital Well-Being from a Confucian Perspective	Dennis & Ziliotti (2023)
Envisioning Communities: A Participatory Approach Towards AI for Social Good	Bondi, Xu, Acosta-Navas, & Killian (2021)
Aligning Artificial Intelligence with Human Values: Reflections from a Phenomenological Perspective	Han, Kelly, Nikou, & Svee (2021)
Challenges of Aligning Artificial Intelligence with Human Values	Sutrop (2020)
The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions	Whittlestone, Nyrup, Alexandrova, & Cave (2019)
Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function)	Eckersley (2019)
Risk Imposition by Artificial Agents: The Moral Proxy Problem	Thoma (2022)

References

Bondi, E., Xu, L., Acosta-Navas, D., & Killian, J. (2021, July). *Envisioning communities: A participatory approach towards AI for social good*. 425–436. https://doi.org/10.1145/3461702.3462612

Christian, B. (2021). *The alignment problem*. New York, NY: WW Norton.

Dennis, M., & Ziliotti, E. (2023). Living well together online: Digital wellbeing from a confucian perspective. *Journal of Applied Philosophy*, *40*(2), 263–279. https://doi.org/10.1111/japp.12627

Eckersley, P. (2019). *Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function)*. arXiv. https://doi.org/10.48550/ARXIV.1901.00064

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Han, S., Kelly, E., Nikou, S., & Svee, E.-O. (2021). Aligning artificial intelligence with human values: Reflections from a phenomenological perspective. *AI & Society*, *37*, 1–13. https://doi.org/10.1007/s00146-021-01247-4

Sutrop, M. (2020). Challenges of aligning artificial intelligence with human values. *Acta Baltica Historiae Et Philosophiae Scientiarum*, *8*(2), 54–72. https://doi.org/10.11590/abhps.2020.2.04

Thoma, J. (2022). Risk imposition by artificial agents: The moral proxy problem. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The cambridge handbook of responsible artificial intelligence: Interdisciplinary perspectives*. Cambridge University Press.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019, January). *The role and limits of principles in AI ethics: Towards a focus on tensions*. 195–200. https://doi.org/10.1145/3306618.3314289