Glossar

Funktionale Architektur. Die abstrakte kausale Struktur eines Systems, die umfasst, wie äußere Einflüsse (Inputs) auf innere Zustände des Systems einwirken, die wiederum andere innere Zustände sowie Reaktionen (Outputs) hervorrufen.

These der instrumentellen Konvergenz. Fast jedes intelligente System hat bestimmte *instrumentelle Werte*. Dazu zählt es, die eigene Existenz und die eigenen intrinsischen Werte zu bewahren, intelligenter zu werden und Ressourcen anzusammeln.

Künstliche Intelligenz (KI). Ein künstlich geschaffenes System, das komplexe Aufgaben bewältigen kann.

KI, allgemeine. KI, deren Fähigkeit, komplexe Aufgaben zu bewältigen, in allen Bereichen der von Menschen zumindest gleichkommt.

KI, enge. KI, die nur in einem eng begrenzten Bereich komplexe Aufgaben bewältigen kann – z. B. ein Schachprogramm.

Künstliches neuronales Netzwerk (KNN). Ein künstliches Netzwerk aus Knotenpunkten – genannt ›(künstliche) Neuronen‹ und Verbindungen zwischen diesen Neuronen. Wenn ein Neuron aktiviert wird, kann sich diese Aktivität über die Verbindungen auf andere Neuronen übertragen, abhängig von der Stärke der Verbindungen zwischen ihnen. KNNs werden üblicherweise trainiert. Sie lernen, bestimmte Aufgaben zu bewältigen, indem sich ihre Verbindungsstärken während des Trainings verändern.

KNN, tiefes. Ein KNN mit mehreren Schichten von Neuronen, von denen einige weder direkt Input von außerhalb des Netzwerks erhalten noch direkt Output nach außen geben.

KNN, vorwärtsgekoppeltes. Ein KNN, in dem die Aktivität von Neuronen nur von hinten nach vorne, d. h. von Input zu Output weitergegeben wird.

Orthogonalitätsthese. Intelligenz und Motivation bzw. Werte sind annähernd voneinander unabhängig. Das heißt, beinahe jeder Grad an Intelligenz ist mit beinahe jedem Wertesystem verträglich.

Phänomenales Bewusstsein. Wie es sich anfühlt bzw. wie es ist, bestimmte Empfindungen zu haben; subjektives Erleben.

Psychologische Kontinuität. Diese liegt genau dann vor, wenn die zeitlich unmittelbar aufeinanderfolgenden geistigen Zustände einer Person einander hinreichend ähnlich sind und kausal voneinander abhängen. Der psychologischen Theorie personaler Identität zufolge hängt unser Fortbestehen von psychologischer Kontinuität ab.

Superintelligenz. Ein System, dessen Intelligenz der von Menschen deutlich überlegen ist.

Vielfache Realisierbarkeit. Bestimmte geistige Zustände, z. B. Freude und Schmerzen, können in verschiedenen Wesen eine andere physiologische Basis haben.

Werte, intrinsische. Die Dinge, die wir unbedingt und um ihrer selbst willen schätzen.

Werte, instrumentelle. Dinge, die wir deshalb schätzen, weil sie uns bei der Erfüllung unserer intrinsischen Werte dienlich sind.

Literatur

Bostrom, Nick: *Superintelligenz. Szenarien einer kommenden Revolution.* Berlin: Suhrkamp 2016.

Brynjolfsson, Erik/McAfee, Andrew: *The Second Machine Age. Wie die nächste digitale Revolution unser aller Leben verändern wird.* Kulmbach: Plassen 2014.

Campbell, Murray/Hoane, Jr., Joseph/Hsu, Feng-hsiung: Deep Blue. In: *Artificial Intelligence* (2002) 134, 57–83.

Chalmers, David: Absent qualia, fading qualia, dancing qualia. In: Thomas Metzinger (Hg.): *Conscious Experience.* Imprint Academic: Thorverton 1995, 309–328.

Dastin, Jeffrey: Amazon scraps secret AI recruiting tool that showed bias against women (2018). *Reuters,* 9. 10. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dou, Eva/Geng, Olivia: AI masters the game of Go (2017). In: *The Wall Street Journal,* 6. 1. 2017.

Frey, Carl Benedikt/Osborne, Michael A.: The future of employment: How susceptible are jobs to computerization? Oxford Martin School, 1. 9. 2013. URL: https://www.oxfordmartin.ox.ac.uk/publications/the-future-of-employment/.

Goodfellow, J. et al.: Explaining and harnessing adversarial examples (2015). arXiv: 1412.6572.

Grace, Katja et al.: When will AI exceed human performance? Evidence from AI experts. In: *Journal of Artificial Intelligence Research* (2018) 62, 729–754.

Human Rights Watch: China: Minority region collects DNA from millions (2017). URL: https://www.hrw.org/news/2017/12/13/china-minority-region-collects-dna-millions. 13. 12. 2017.

KI-Expertengruppe: Ethik-Leitlinien für eine vertrauenswürdige KI. Hochrangige Expertengruppe für Künstliche Intelligenz, eingesetzt von der Europäischen Kommission im Juni 2018 (2019). URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60425.

Kurzweil, Ray: *Menschheit 2.0: Die Singularität naht.* Berlin: Lola Books 2014.

Misselhorn, Catrin: *Grundfragen der Maschinenethik.* Ditzingen: Reclam 2018.

Nebehay, Stephanie: U.N. says it has credible reports that China holds million Uighurs in secret camps (2018). *Reuters,* 10. 8. 2018. URL: https://www.reuters.com/article/us-china-rights-un/u-n-says-it-has-credible-reports-that-china-holds-million-uighurs-in-secret-camps-idUSKBN1KV1SU.

O'Neil, Cathy: *Angriff der Algorithmen: Wie sie Wahlen manipulieren, Berufschancen zerstören und unsere Gesundheit gefährden.* München: Hanser 2017.

Rajpurkar et al.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning (2017). *arXiv:* 1711.05225.

Silver, David et al.: Mastering the game of Go without human knowledge (2017). In: *Nature.* DOI: 10.1038/nature24270.

Silver, David et al.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. In: *Science* 362 (6419) (2018), 1140–1144.

Simmons, Keir: Inside Chinese camps thought to be detaining a million Muslim Uighurs (2019). *NBC News,* 4. 10. 2019. URL: https://www.nbcnews.com/news/world/inside-chinese-camps-thought-detain-million-muslim-uighurs-n1062321.

Slaughterbots (2017). *YouTube.* 14. 12. 2017. URL: https://www.youtube.com/watch?v=9CO6M2HsoIA.

Tegmark, Max: *Leben 3.0. Mensch sein im Zeitalter Künstlicher Intelligenz.* Berlin: Ullstein 2017.

The complicated truth about China's social credit system (7. 6. 2019). *Wired.*

Vinyals, Oriol et al.: StarCraft II. A new challenge for reinforcement learning (2017). *arXiv:* 1708.04782.

Vinyals, Oriol et al.: Grandmaster level in StarCraft II using multi-agent reinforcement learning (2019). In: *Nature.* DOI: 10.1038/s41586-019-1724-z.

Walsh, Toby: Expert and non-expert opinion about technological unemployment (2017). *arXiv:* 1706.06906.

Wilson, William J.: *When Work Disappears. The World of the New Urban Poor.* New York: Knopf 1997.
World Inequality Lab: *Bericht zur weltweiten Ungleichheit* (2018).
Yudkowsky, Eliezer: Staring at the singularity (1996). URL: http://yudkowsky.net/obsolete/singularity.html.