References & Recommended Reading

Sam Altman, AI's biggest star, sure hopes someone figures out how not to destroy humanity
https://www.cnn.com/2024/12/05/business/sam-altman-openai-nightcap

IASEAI Issues Call to Action for Lawmakers, Academics, and the Public Ahead of AI Summit in Paris
https://www.iaseai.org/conference/statement

AI vs. machine learning vs. deep learning vs. neural networks: What's the difference?
https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

Nexus (book) - Yuval Noah Harari (If you read one non-fiction book this year, read this one!)

AI alignment - Wikipedia
https://en.wikipedia.org/wiki/AI_alignment

The Memory Paradox: Why Our Brains Need Knowledge in an Age of AI (2025-05-11)
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5250447

The Alignment Problem (book) - Brian Christian
Obama quote - page 296

Flynn effect
https://en.wikipedia.org/wiki/Flynn_effect

4 points from FB whistleblower (Frances Haugen) testimony on Capitol Hill (2021-10-05)
https://www.npr.org/2021/10/05/1043377310/facebook-whistleblower-frances-haugen-congress

10-year ban on US states regulating AI in BBB
https://apnews.com/article/ai-regulation-state-moratorium-congress-39d1c8a0758ffe0242283bb82f66d51a

Why Colorado's artificial intelligence law is a big deal for the whole country
https://coloradosun.com/2025/04/23/colorado-artificial-intelligence-law-ai/

Reward hacking - Wikipedia
https://en.wikipedia.org/wiki/Reward_hacking

Blake Lemoine: Google fires engineer who said AI tech has feelings (BBC)
https://www.bbc.com/news/technology-62275326

Meta sues maker of explicit deepfake app for dodging it rules to advertise AI 'nudifying' tech
https://www.cnn.com/2025/06/12/tech/meta-sues-explicit-deepfake-app-crushai

Introducing Superalignment
https://openai.com/index/introducing-superalignment/

GPT-4 Technical Report (7 Mar 2023, Description of ARC TaskRabbit test)
https://cdn.openai.com/papers/gpt-4.pdf (see page 55)

OpenAI's GPT-4 faked being blind to deceive a TaskRabbit human into helping it solve a CAPTCHA
https://www.foxbusiness.com/technology/openais-gpt-4-faked-being-blind-deceive-taskrabbit-human-helping-solve-captcha

Nosedive (Black Mirror (Netflix series), 2016)
https://en.wikipedia.org/wiki/Nosedive_(Black_Mirror)

Sky wars: the race for drone dominance (June 2025)
https://rationaloptimistsociety.substack.com/p/sky-wars-the-race-for-drone-dominance

Our AI Future Is WAY WORSE Than You Think | Yuval Noah Harari (October 2024)
https://www.youtube.com/watch?v=_jl64f-821o

U.S. Enacts Take It Down Act (June 2025)
https://natlawreview.com/article/us-enacts-take-it-down-act

Largest AI companies by market capitalization
https://companiesmarketcap.com/artificial-intelligence/largest-ai-companies-by-marketcap/

Weapons of Math Destruction - Cathy O’Neil