Ricercatori e intellettuali: paper e riferimenti chiave

Riferimenti organizzati per figura, con link ai paper principali
Compilato il 27 febbraio 2026

---

1. Dario Amodei

Google Scholar: scholar.google.com/citations?user=6-e-ZBEAAAAJ | Semantic Scholar: semanticscholar.org/author/2698777 | Saggi: darioamodei.com

Fase accademica (biofisica, Princeton/Stanford)

• **ProteoWizard: a unified data access interface for proteomics.** *Bioinformatics* (2012).
Toolkit open-source per la ricerca proteomica — dal periodo post-doc a Stanford con Parag Mallick.

Fase Baidu (2014–2015)

• Amodei, D. et al., **Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.** *ICML 2016*.
Sistema di riconoscimento vocale end-to-end; MIT Technology Review top 10 breakthrough 2016.
→ arxiv.org/abs/1512.02595

Fase OpenAI (2016–2021)

• Amodei, D., Olah, C., Steinhardt, J. et al., **Concrete Problems in AI Safety.** *arXiv* (2016).
Paper fondativo sulla safety AI: 5 problemi concreti di rischio da incidenti nei sistemi di ML.
→ arxiv.org/abs/1606.06565

• Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., **Language Models are Unsupervised Multitask Learners.** (2019).
Il paper di GPT-2.
→ cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

• Brown, T.B. et al. (+30 autori), **Language Models are Few-Shot Learners.** *NeurIPS 2020*.
Il paper di GPT-3. Amodei ultimo autore (senior).
→ arxiv.org/abs/2005.14165

• Kaplan, J., McCandlish, S. et al., **Scaling Laws for Neural Language Models.** *arXiv* (2020).
Paper fondamentale sulle leggi di scaling — più grande il modello e i dati, migliori le performance in modo prevedibile.
→ arxiv.org/abs/2001.08361

• Ziegler, D.M. et al., **Fine-Tuning Language Models from Human Preferences.** *arXiv* (2019).
Uno dei paper fondativi del RLHF (Reinforcement Learning from Human Feedback).
→ arxiv.org/abs/1909.08593

• Stiennon, N. et al., **Learning to Summarize from Human Feedback.** *NeurIPS 2020*.
Applicazione del RLHF alla summarization.
→ arxiv.org/abs/2009.01325

Fase Anthropic (2021–)

• Bai, Y. et al., **Training a Helpful and Harmless Assistant with RLHF.** *arXiv* (2022).
Paper core dell'approccio Anthropic: come bilanciare helpfulness e harmlessness.
→ arxiv.org/abs/2204.05862

• Bai, Y. et al., **Constitutional AI: Harmlessness from AI Feedback.** *arXiv* (2022).

Introduce il Constitutional AI — il sistema si auto-corregge sulla base di principi scritti, senza bisogno costante di feedback umano.

→ arxiv.org/abs/2212.08073

• Ganguli, D. et al., **Red Teaming Language Models to Reduce Harms.** *arXiv* (2022).

Metodologie di red teaming per identificare comportamenti dannosi.

→ arxiv.org/abs/2209.07858

• Ganguli, D. et al., **The Capacity for Moral Self-Correction in Large Language Models.** *arXiv* (2023).

I LLM possono auto-correggersi moralmente se istruiti a farlo.

→ arxiv.org/abs/2302.07459

• Elhage, N. et al., **Toy Models of Superposition.** *arXiv* (2022).

Paper di interpretabilità: come i modelli rappresentano più feature di quante abbiano dimensioni.

→ arxiv.org/abs/2209.10652

Saggi pubblici

• Amodei, D., **Machines of Loving Grace.** (2024).

Saggio lungo sulle possibilità positive dell'AI.

→ darioamodei.com/machines-of-loving-grace

2. Demis Hassabis

Google Scholar: scholar.google.com/citations?user=dYpPMQEAAAAJ | h-index: 83 | Citazioni: 262.000+ | Nobel Chimica 2024

Fase neuroscientifica (UCL, 2005–2009)

• Hassabis, D. et al., **Patients with hippocampal amnesia cannot imagine new experiences.** *PNAS* 104(5):1726-31 (2007).

Paper rivoluzionario: i pazienti con amnesia ippocampale non riescono nemmeno a immaginare nuove esperienze. Stabilisce il legame tra memoria episodica e immaginazione. Science top 10 scoperte 2007.

• Hassabis, D. et al., **Using imagination to understand the neural basis of episodic memory.** *J. Neuroscience* 27(52):14365-74 (2007).

DeepMind — Reinforcement Learning e giochi

• Mnih, V. et al., **Human-Level Control through Deep Reinforcement Learning.** *Nature* 518:529-33 (2015).

L'agente che impara a giocare ai giochi Atari a livello umano. Paper fondativo del deep reinforcement learning.

→ nature.com/articles/nature14236

• Silver, D. et al., **Mastering the Game of Go with Deep Neural Networks and Tree Search.** *Nature* 529:484-489 (2016).

AlphaGo batte il campione mondiale di Go.

→ nature.com/articles/nature16961

• Silver, D. et al., **Mastering the Game of Go without Human Knowledge.** *Nature* 550:354-359 (2017).

AlphaGo Zero: impara da zero, senza dati umani.

→ nature.com/articles/nature24270

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

AlphaZero: un singolo algoritmo che padroneggia tre giochi.

DeepMind — Scienza

• Jumper, J. et al., **Highly accurate protein structure prediction with AlphaFold.** *Nature* 596:583-589 (2021).

AlphaFold2. Nobel per la Chimica 2024. Breakthrough of the Year 2021 (Science). Uno dei 10 paper più citati degli ultimi 5 anni.

→ nature.com/articles/s41586-021-03819-2

• Davies, A. et al., **Advancing mathematics by guiding human intuition with AI.** *Nature* (2021).

AI usata per guidare l'intuizione matematica umana.

→ nature.com/articles/s41586-021-04086-x

Paper teorico chiave

• Hassabis, D. et al., **Neuroscience-Inspired Artificial Intelligence.** *Neuron* 95(2):245-58 (2017).

Paper programmatico: capire meglio i cervelli biologici è essenziale per costruire macchine intelligenti.

→ pubmed.ncbi.nlm.nih.gov/28728020/

Altro

• Graves, A. et al., **Hybrid Computing Using a Neural Network with Dynamic External Memory.** *Nature* 538:471-76 (2016).

Il Differentiable Neural Computer — rete neurale con memoria esterna.

• Kirkpatrick, J. et al., **Overcoming Catastrophic Forgetting in Neural Networks.** *PNAS* 114(13):3521-26 (2017).

Come prevenire il “forgetting catastrofico” nelle reti neurali.

3. Geoffrey Hinton

Google Scholar: scholar.google.com/citations?user=JicYPdAAAAAJ | Lista completa: cs.toronto.edu/~hinton/papers.html | Citazioni: 1.009.000+ | Nobel Fisica 2024 | Turing Award 2018

I paper fondativi

• Rumelhart, D.E., Hinton, G.E., Williams, R.J., **Learning representations by back-propagating errors.** *Nature* 323:533-536 (1986).

IL paper della backpropagation. Ha reso possibile l'addestramento delle reti neurali profonde. Probabilmente il paper più influente nella storia dell'AI.

→ nature.com/articles/323533a0

• Ackley, D.H., Hinton, G.E., Sejnowski, T.J., **A Learning Algorithm for Boltzmann Machines.** *Cognitive Science* 9:147-169 (1985).

Introduce le macchine di Boltzmann — fondamentali per l'apprendimento non supervisionato.

• Hinton, G.E., Salakhutdinov, R.R., **Reducing the Dimensionality of Data with Neural Networks.** *Science* 313:504-507 (2006).

Paper che ha rilanciato il deep learning dopo il “winter” degli anni '90. Introduce le Deep Belief Networks.

La rivoluzione del deep learning

• Krizhevsky, A., Sutskever, I., Hinton, G.E., **ImageNet Classification with Deep Convolutional Neural Networks.** *NIPS 2012*.

AlexNet. 126.000+ citazioni. Ha dimostrato definitivamente la superiorità del deep learning nella classificazione di immagini.

• Srivastava, N., Hinton, G. et al., **Dropout: A Simple Way to Prevent Neural Networks from Overfitting.** *JMLR* 15:1929-1958 (2014).

Introduce il dropout — tecnica usata praticamente ovunque nel deep learning moderno.

• Hinton, G., Vinyals, O., Dean, J., **Distilling the Knowledge in a Neural Network.** *arXiv* (2015).

Knowledge distillation — comprimere la conoscenza di un modello grande in uno piccolo.

→ arxiv.org/abs/1503.02531

Review e visione

• LeCun, Y., Bengio, Y., Hinton, G., **Deep Learning.** *Nature* 521:436-444 (2015).

LA review definitiva sul deep learning, scritta dai tre “padri” del campo. 60.000+ citazioni.

→ nature.com/articles/nature14539

• Sabour, S., Frosst, N., Hinton, G.E., **Dynamic Routing Between Capsules.** *NIPS 2017*.

Introduce le Capsule Networks — tentativo di superare i limiti delle CNN.

→ arxiv.org/abs/1710.09829

• van der Maaten, L., Hinton, G., **Visualizing Data using t-SNE.** *JMLR* (2008).

Algoritmo t-SNE per la visualizzazione di dati ad alta dimensionalità.

Nota biografica: Hinton lasciò gli USA per il Canada nel 1987 per motivi politici, opponendosi al finanziamento militare della ricerca AI sotto Reagan. Parallelo diretto con la vicenda Amodei-Pentagono di oggi.

4. Lina Khan

Non una ricercatrice scientifica ma una giurista-accademica. Un solo paper ha ridefinito un campo intero.

• Khan, L.M., **Amazon's Antitrust Paradox.** *Yale Law Journal* 126:710-805 (2017).

Scritto da studentessa di legge a Yale. Argomenta che il framework antitrust basato sul "consumer welfare" (= prezzi bassi = nessun problema) non funziona nell'era delle piattaforme digitali. Il NYT lo definì una "rifrazione di decenni di legislazione sul monopolio."

→ PDF: economics.utah.edu/antitrust-conference/Amazons%20Antitrust%20Paradox.pdf
→ Columbia Law: scholarship.law.columbia.edu/faculty_scholarship/2808/
→ SSRN: papers.ssrn.com/sol3/papers.cfm?abstract_id=2911742
→ Yale: openyls.law.yale.edu/handle/20.500.13051/10275

Khan ha anche scritto la tesi di laurea al Williams College su Hannah Arendt.

---

5. Aaron Swartz (1986–2013)

Programmatore, attivista, scrittore. Non un accademico in senso formale: co-sviluppò RSS a 14 anni, contribuì a Creative Commons, co-fondò Reddit. Morto suicida a 26 anni sotto il peso di un'incriminazione federale per aver scaricato articoli accademici da JSTOR.

Scritto principale

• Swartz, A., **Guerilla Open Access Manifesto.** Eremo, Italia, luglio 2008.

Manifesto per la liberazione della conoscenza scientifica dai paywall. Pubblico dominio. Sci-Hub di Alexandra Elbakyan (74M+ articoli gratis) è esplicitamente ispirato a questo manifesto.

→ archive.org/details/GuerillaOpenAccessManifesto

Contributi tecnici

RSS 1.0 (co-autore della specifica, 2000 — aveva 14 anni) • Creative Commons (2002–) • web.py (framework web Python) • Reddit (co-fondatore, 2005) • Open Library • Liberazione dei documenti PACER (2008)

Letteratura secondaria

• Swift, K., **A Web of Extended Metaphors in the Guerilla Open Access Manifesto of Aaron Swartz.** UC Santa Barbara, 2019.

Analisi linguistica e di frame semantici del Manifesto.

→ escholarship.org/uc/item/6w76f8x7

Documentario: "The Internet's Own Boy: The Story of Aaron Swartz" (2014), di Brian Knappenberger. Disponibile gratuitamente su Internet Archive.

---

6. Riferimenti incrociati e storici

J. Robert Oppenheimer

• Bird, K. & Sherwin, M.J., **American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer.** (2005).

La biografia definitiva, Premio Pulitzer.

Andrei Sakharov

• Sakharov, A.D., **Reflections on Progress, Peaceful Coexistence, and Intellectual Freedom.** (1968).

Il saggio che lo rese dissidente: argomenta per la convergenza tra sistemi, la libertà intellettuale, e il disarmo nucleare. Scritto dal padre della bomba H sovietica.

→ sakharov-archive.ru

Documento compilato il 27 febbraio 2026 a partire da Google Scholar, Semantic Scholar, arXiv, PubMed, DBLP, e fonti primarie.