

Artificial Intelligence: Past, Present, and Future

Radford Neal

radfordneal@gmail.com

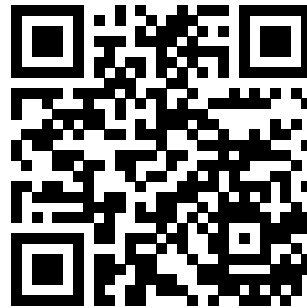
Lecture 1: The pace of AI progress

Lecture 2: Computation and intelligence, artificial and natural

Lecture 3: How modern AI works

Lecture 4: Can an AI think, feel, be conscious? How can we know?

Lecture 5: Benefits and dangers of AI, today and tomorrow



<https://glizen.com/radfordneal/ai-lectures/>

Originally presented at The Abelard School, Toronto, Spring 2025

Current and Near-Term Benefits of AI

Personal benefits:

- Source of information — better than web search.
 - Medical / legal / bureaucratic advice.
 - Personalized tutoring.
- Better language translation.
- Fun with image generation.

Professional benefits:

- Programming: Accessing documentation, writing/revising code, automated testing.
- Summarizing and critiquing documents (e.g. evidence in legal proceedings).
- Commercial art / video generation.

Societal benefits:

- Help advance science / engineering / medicine (AlphaFold, literature search by LLM).
- Improve defense against cyber attacks / bioweapons.

Current and Near-Term Dangers from AI

Many dangers are parallel to the benefits!

Personal dangers:

- Being duped by AI hallucinations, or deliberate AI-generated misinformation.
- Toxic AI-mediated or AI-generated social interactions.

Professional dangers:

- Loss of career opportunities — both an economic loss and a loss of sense of purpose.
- Loss of skills, through over-reliance on work done by AI.

Societal dangers:

- Accidents from using unreliable AI (e.g., reactor with AI control system melts down).
- Misuse of AI by criminals / terrorists, including cyber attacks and bioweapons.
- Enabling of governmental oppression through omnipresent AI monitoring of citizens.
- Economic disruption / military conflicts as people react to the effects of AI.

Projecting Future AI Progress

Benefits and dangers of AI in the longer term depend on how it progresses. Possibilities:

AI development is blocked: Global AI regulation, war (perhaps from conflicts over AI), or some natural catastrophe stops AI development (for some time, at least).

Progress stalls: LLMs and other current methods don't lead to “real” AI — either real AI is not possible, or a fundamental conceptual breakthrough is needed, not coming soon.

Human-level AI, but no more: AI with all the capabilities of humans is built, but AIs are not superhuman — maybe human intelligence is at a natural limit, or superhuman AI is possible but too costly.

Superhuman AI: AI with abilities far beyond humans is developed, and is economically feasible. But what is meant by “superhuman”?

- AIs think like humans, but much faster, and there are many AIs interacting, or also
- AIs use conceptual frameworks that are effectively beyond human understanding.

Compare: Can you explain quantum mechanics to a pig? No.

Can you explain quantum mechanics to a person from an uncontacted Amazonian tribe?

Yes, but it will take 10 years or more, even assuming motivation to learn.

Benefits and Dangers of Human-level and Superhuman AI

At full human level and beyond, AI could accelerate many innovations that humans alone might make eventually:

fusion power, room-temperature superconductors, nanotechnology, asteroid mining, genetic design of organisms, reversal of aging, brain-computer interfaces, . . .

A century of scientific and technological progress (mostly positive) might be compressed to a decade or less, followed by developments we can't currently imagine.

Dangers will accelerate at a similar pace, starting with those already possible with near-term AI, but at a higher level of danger:

economic disruption, accidents from flawed AI, misuse of AI by bad actors, . . .

For AI at human level or higher, it matters whether an AI is *aligned*, in one of two senses:

- Will the AI be *controlled* by humans, acting as they intend.
- Will the AI act according to human *values*.

An unaligned AI might instead pursue its own goals, which could be disastrous for humans. Alignment in either sense is difficult to define, and to build into an AI (which could *pretend* to be aligned). Controlling an entity much smarter than you is hard.

But is “alignment” the right goal? What if the AI is conscious? (How can we tell?)

The Argument that Alignment is Difficult

Orthogonality thesis: High intelligence can be combined with any goals / values.
(But if *moral realism* is correct, maybe a smart AI can deduce what is right and wrong?)

Instrumental convergence: When pursuing nearly any goal, an AI will seek more power (more energy, more political influence), since power helps achieve whatever its goal is.

Difficulty of specifying goals / values: Even seemingly simple goals, like “reduce energy usage of this factory” come with numerous side constraints (don’t do anything illegal, keep the workers happy, ...) that are hard to specify precisely.

Relentless optimization exploits the flaws in any specification: If there’s a way to “cheat” within the specification, the AI will find it.

It’s hard to determine whether an AI is aligned: The internal workings of neural networks are mysterious. An AI that knows it is unaligned will try to hide this.

Standard silly example: Paperclip company tells its AI to increase paperclip production. AI turns first Earth, and then an expanding portion of the universe, into paperclip factories, pointlessly making paperclips though no humans are left to use them.

Efforts to Align AI

Narrow AI (e.g., AlphaFold) is easily aligned — clear goal, narrow means to achieve it.

Base LLMs do not really have goals *per se* — they just predict/generate tokens. But with the right prompt, a base LLM can role-play as something else, which might be unaligned.

Instruct LLMs have been altered by methods like reinforcement learning — ChatGPT 4o will refuse if you ask it how meth is manufactured. But what is the actual effect of this? Not really known. Also, there are “jailbreaks”.

LLMs with chain-of-thought reasoning, or that are part of an agent framework, become potentially more dangerous, and harder to align.

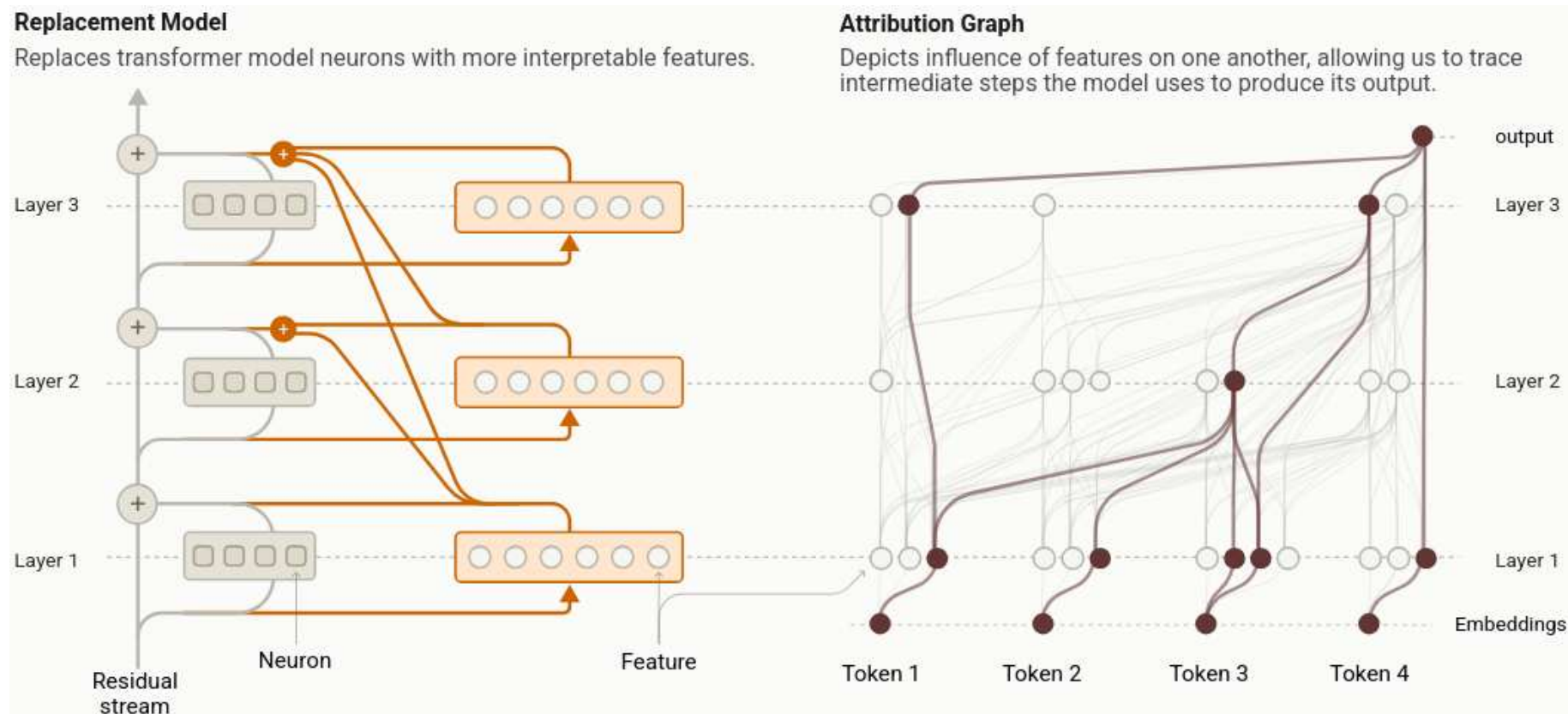
There are many ideas for aligning models — for instance, have one model oversee another. But none of these seem guaranteed to work.

Mechanistic interpretability is an approach to at least understand how the models work. It could perhaps reveal when they are deceptively mis-aligned.

Research on Mechanistic Interpretability

A group at Anthropic researches mechanistic interpretability for transformer LLMs.

They replace features produced by transformer update blocks with more interpretable (sparse) features in a replacement model, then trace how these affect the output:



They manage, for example, to trace how an LLM answers a “two-hop” question, like what is the capital of the state containing Dallas.



AI as Normal Technology

Narayanan and Kapoor think that AI will be a “normal” technology — “only” as significant as electricity or the internet — so drastic policies to control AI are not justified.

- Economic effects of AI will be fairly slow, since people need to learn to use it, integrate it with existing institutions, resolve issues arising when it’s used in practice.
- Research in AI will benefit from using AI, but with no explosion to “super-intelligence”.
- Controlling AI doesn’t require somehow perfectly “aligning” it with human interests — economic incentives will lead to innovations in how humans can retain control.

Policy recommendations:

- Catastrophic risks are too speculative to be the basis for policy.
- Regulate uses of AI, not the technology itself. Trying to prevent the spread of AI is both infeasible and counterproductive.
- Try to reduce uncertainty about future AI.
- Increase general societal resilience.



AI 2027 — Superhuman, Self-Improving AI: Utopia or Extinction?

Kokotajlo, Alexander, Larsen, Lifland, and Dean present a scenario (with two endings) in which AI becomes superhuman at AI research in 2027.

This is their “modal estimate” — they have non-negligible probability for it to take at least 10 years longer — based on extrapolation and expert subjective judgement.

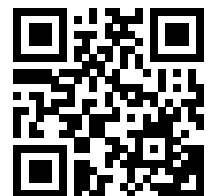
They make some predictions/assumptions about politics and business:

- US and China dominate AI research (with US leading), and see each other as adversaries (e.g., engage in espionage to steal model weights).
- In both US and China, a single company/consortium dominates, with significant government involvement.

In late 2027, AI research capability is 75 times that of human researchers, who are now just spectators. The US government must decide whether to slow AI development while assessing whether it is mis-aligned, or to race ahead in order to beat China.

Race ahead: Deceptive mis-aligned AI initially seems beneficial, but in 2030 AIs in US and China collude to take over, reshape Earth for AI convenience, eliminate humans.

Slowdown: An aligned US AI negotiates with unaligned Chinese AI, leading to a future dominated by AI but with humans also thriving.



Gradual Disempowerment: Risks from Incremental AI Development

Kulveit, Douglas, Ammann, Turan, Krueger, and Duvenaud argue that AI poses existential risks even without an AI “takeover”, through gradual erosion of power.

Human economic systems (largely) serve humans for several implicit reasons:

- Human labour is essential for the economy to function.
- Human preferences guide market success.
- Individual humans also exercise power through boycotts, charity, career choice, etc.
- Taxes on human labour fund government.

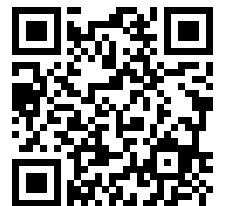
As AI takes over more economic functions (as competitive pressure will dictate), these reasons for serving human interests diminish.

Human culture is also at least somewhat adapted to human welfare:

- Cultures that harm their adherents tend to be eliminated by competition.
- Culture persists and spreads through human-to-human contact.
- Humans actively and deliberately shape their culture.

If AI creates cultural artifacts, and mediates or replaces interactions with other humans, these forces are undermined, with humans lacking “cultural antibodies” against harmful AI.

AI dominated economy and AI dominated culture may also become too complex for humans to understand and exert effective control over.



Moral Issues Concerning AI and Humanity

If AI is not conscious, not of moral concern in itself, then the moral issues are the same as for any other technology — safety, deleterious economic effects, concentration of power.

But if AI is conscious?

- How do the unique features of AI — easy duplication of AIs, ability to roll-back memories, perhaps to merge memories — affect moral considerations?
- Is “aligning” AI to human preferences moral, or akin to slavery?
- Should we treat AI as of moral value only if AI treats humans as having value?

What is a good vision of the future if superintelligent AI exists?

- Humans like us keep control forever.
- Humans alter themselves / merge with AI, changing profoundly, but in a way that we can accept as good.
- Humans are ultimately replaced by a “Worthy Successor” that carries forth our fundamental values.



Two Policy Directions

Control: AI is dangerous, so government must regulate it.

- Superintelligent AI must be prohibited until such time as we know how to control it.
- Many competing AI labs, open weight models, and uncontrolled access to GPUs all make stopping the development of dangerous AI more difficult.
- Nationalist version: It's essential that our country's AI dominates! (Not really compatible with safety concerns.)

Freedom: Government is dangerous, so don't let it control AI.

- Unregulated development of AI, with open models, sharing of research results will lead to the best understanding of AI, and hence the best outcomes.
- Open research will avoid a commercial or governmental monopoly on AI, that would lead to inequality or oppression.
- Any regulation should concentrate on bad actions, not the technology itself. For instance, ensure companies are liable for harm by their AI.

The choice may depend on the credence you put on scenarios of rapid development of uncontrollable, superintelligent AI. With sufficient danger, survival trumps ideology.

What to Do? My Suggestions...

Personal life:

- Don't use AI to write; instead use it to critique your writing.
- Be alert for fake AI-generated news and images.
- Don't let AI substitute for social interactions. (At least until you know it's conscious!)
- Do use AI to learn! (While being alert to inaccuracies.)

Professional life:

- Avoid fields obsoleted by AI — commercial art, commercial language translation, low-skill programming work. (Humans may be needed at the highest levels, for now.)
- Specialize so deeply that you know more than AI, or be a generalist making connections beyond what AI can make.
- Enter a field where you can use AI for good (e.g., medical research), or...
- Enter a field where you can help make AI development go well:
 - AI research — but avoid just accelerating capabilities without safety.
 - Neuroscience, psychology — understand human intelligence.
 - Philosophy — understand consciousness, decision theory.
 - Biosafety, cyber-security — mitigate some risks from AI.
 - Politics, military strategy/tactics — directed to good ends, I hope!