Science & technology

Artificial intelligence

Computo, ergo sum

Why big AI labs are hiring philosophers to shape their products

TEN YEARS ago, as the AI revolution was gathering pace, arts and humanities students were told that, if they wanted to make themselves employable, they should "learn to code". That may have been bad advice. These days, it is programmers who are nervous about AI taking their jobs.

They might consider learning to philosophise. Earlier this year the Federal Reserve Bank of New York published figures showing that American philosophy graduates are more likely to have jobs than their peers who studied computer science. In 2024, the most recent year for which numbers are available, 7% of those who had studied computer science were unemployed, against just 5.1% of philosophers.

Many are being snapped up by AI firms themselves. Students get job offers before they have graduated, says Luciano Floridi, a philosopher at Yale University. Academics are moving, too. Dr Floridi describes the scale of departures from philosophy departments as a "haemorrhaging".

Some of the lessons that philosophy can offer AI researchers are ancient. The Socratic method—as described by Plato, an ancient Greek philosopher—uses feigned ignorance and sequential questioning to clarify meanings, spot contradictions and reveal ramifications. Many current AI systems tend towards sycophancy. Models trained in the Socratic method, says Jörg Noller, an expert on philosophy and AI at Ludwig Maximilian University of Munich, are less keen on people-pleasing and more willing to pursue the truth.

Then there is the idea of "Socratic ignorance". In the "Apology", Plato has Socrates claim that his wisdom consists mostly of being aware of how much he does not know. Implanting that humility into a model can help limit overconfidence, a common flaw that Dr Noller describes as "AI immaturity". Iason Gabriel, a senior philosopher at Google DeepMind, an AI lab based in London, attributes an industry-wide decline in hallucinations to such efforts. More broadly, he says, philosophy lessons are "a powerful mechanism" for improving long AI reasoning processes known as "chains of thought".

Philosophical training can also affect a model's outlook in more specific ways. Feed an AI legal assistant the writings of John Locke, says Thomas Powers, a philosopher of technology at the University of Delaware, and it will favour robust property rights as an underpinning of political liberty. And if you don't like those principles, the model-makers have others. The "Granite" series of models from IBM, an American computing giant, come with dials that let business customers better align outputs with their own corporate philosophies. Francesca Rossi, IBM's head of responsible AI, says these can let users choose where to strike the balance between philosophical trade-offs, such individual agency versus social harmony.

Philosophy can help with safety, too. Researchers have documented all sorts of ominous behaviour in AI models, including attempts to evade oversight and even blackmail their users. One way model-makers try to discourage this sort of misbehaviour is called AI constitutionalism. This involves building a model around a scaffolding of rules and principles culled from philosophical writings with legal or moral authority.

Anthropic, an AI lab based in San Francisco, is one proponent. Constitutions for its Claude models have incorporated material from sources as diverse as Immanuel Kant, Apple's terms of service and the Universal Declaration of Human Rights. The latest iteration, led by Anthropic's top philosopher, Amanda Askell, was published on January 21st. Some staff at Anthropic have nicknamed the 78-page constitution Claude's "soul doc".

The biggest question, though, is what sorts of rules should be put in those constitutions in the first place. Philosophers have zeroed in on two main ethical frameworks. One is deontology. Popular with Kant, among others, this imposes strict rules that prohibit things like lying, coercion and treating people as a means rather than an end, even if it is for a greater good. Anthropic's constitution incorporates many deontological strictures. These can make AI behaviour more consistent, says Dr Powers—a plus for deploying robots in homes and public spaces.

Models with a deontological take on the world have other benefits. One is greater honesty, a trait widely noted in Claude. Models that are more truthful, says Nick Bostrom, a philosopher at the University of Oxford, are less likely to mislead their users. Inflection AI, another Silicon Valley lab, imposes deontological constraints onto its Pi chatbot, which is designed to provide emotional support. Sean White, its boss, says Pi is good at spotting users at risk of harming themselves or others. Deontological constitutions also help with legal compliance, says Dr Floridi.

The other approach to ethics of interest to philosophers of AI is called consequentialism. It weighs costs against benefits to decide what to do. Models more sympathetic to consequentialism include OpenAI's ChatGPT and Google's Gemini. Google's AI models are designed to produce "likely overall benefits [that] substantially outweigh the foreseeable risks", a classic consequentialist goal.

Consequentialist algorithms are also crucial in software for autonomous vehicles: if an accident is unavoidable, a decision must be made on the least tragic way to crash. Chris Gerdes, a senior engineer at Waymo, which makes self-driving cars, says the trend is to make driving software more consequentialist. Consequentialism is also central to AI weapon systems. Military objectives must be weighed against possible civilian deaths, says Jack Shanahan, a former head of the Joint Artificial Intelligence Centre, which studies AI for America's armed forces.

Thorny problems abound—a philosopher's favourite sort. Are there cases when deontological rules should be overridden? How do you make decisions when the consequences are unclear? Should AI systems take into account animal welfare, or the state of the environment? Would it be morally acceptable, asks Stefan Heck, a philosopher and the boss of Nauto, which makes AI-powered safety systems for lorries and other commercial vehicles, to prioritise young pedestrians over old ones? He predicts ethically fraught lawsuits: consequentialist algorithms, after all, explicitly permit one harm as long as it is designed to avert a worse one.

Critics fret about "moral deskilling": if computers increasingly make ethical calls, might people become less willing to make their own judgments? Roman Yampolskiy, an AI theoretician at the University of Louisville, argues that morality "is historically unstable, culturally variable, strategically manipulable, and often only retrospectively legible". Unemployed coders take note: there seems to be no shortage of work for philosophers of AI. ■

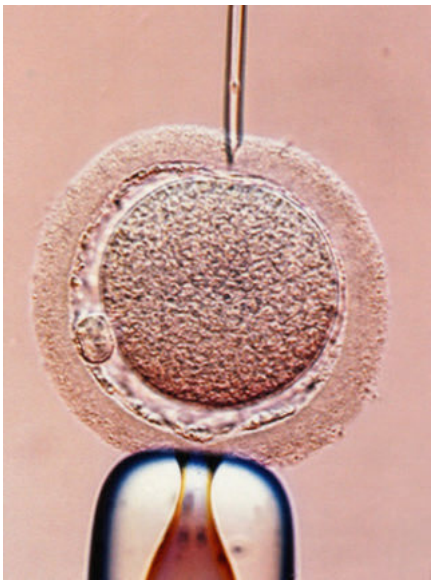
Fertility

Misconceptions

A review of popular "IVF add-on" procedures suggest most do little to help

TRYING TO CONCEIVE through in vitro fertilisation (IVF) leads to disappointment more often than not. About 60% of IVF attempts fail. Fertility clinics offer a long list of tests and procedures purported to boost their customers' chances. But a review of the evidence, published on June 23rd in the *Lancet Obstetrics, Gynaecology & Women's Health*, found no convincing evidence that most of these "IVF add-ons" are helpful. Worryingly, lots of the published evidence was dubious.

The review examined 157 randomised trials of various IVF add-ons. The researchers ran each through a checklist designed to spot signs that the data may have been manipulated. That list was developed in 2023 by fertility researchers (including some of the review's authors) who had noticed a growing number of fraudulent studies.They looked for things such as implausible study timelines, strange participant data or holes in a trial's paper trail.



40% of the time, it works every time

All told, nearly half of the trials did not pass muster. The remaining 85 covered ten commonly used add-ons. Of those, only three procedures showed evidence of abenefit, although the evidence was not particularly strong. The three procedures in question were endometrial scratching (which involves deliberately disturbing the lining of the uterus), EmbryoGlue (in which an embryo is dipped in a solution of hyaluronic acid, which is naturally found in the reproductive tract), and physiological intracytoplasmic sperm injection or PICSI (which tries to select high-quality individual sperm cells).

The data on three other IVF add-ons, though also limited, suggested they had no effect on the chances of a successful pregnancy. These three were corticosteroids (a class of anti-inflammatory drugs), genetic testing of the embryo for abnormal chromosomes and biopsy of the uterine lining to assess genetic expression.

Data for the remaining four procedures were too scant to make a judgment either way. These were acupuncture, intravenous infusion of fats derived from eggs or soyabeans (in the hope this might calm an immune reaction to the embryo), and injections of platelet-rich plasma into either the uterus or the ovaries (platelets being rich in tissue-rejuvenating proteins).

The trials were not just few in number but mostly small in scale too. The median trial had only about 160 patients, which would limit its ability to detect small or subtle effects. The reviewers could not rule out the idea that some interventions might work for a subset of patients. But being able to say who, if anyone, might benefit from any of them would require more and better research—as well as eagle-eyed journal editors weeding out the fishy sort before it is published. ■