



Kent Academic Repository

Mbioh, Will (2026) *Capability determinism, energy, and AI labour substitution*.
AI & Society Journal of Human-Centred Systems and Machine Intelligence
. ISSN 0951-5666.

Downloaded from

<https://kar.kent.ac.uk/115751/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1007/s00146-026-03251-y>

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site.
Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal**, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).



Capability determinism, energy and AI labour substitution

Will Mbioh¹

Received: 10 May 2026 / Accepted: 8 July 2026
© The Author(s) 2026

Abstract

Contemporary policy and media discourse on AI and labour follows a recurring pattern. A demonstration of model capability runs through an inference about workplace substitution to a distributional politics, with the variables that would decide the outcome treated as background. This paper names that pattern capability determinism situates it within Wyatt's account of soft technological determinism, and reframes the substitution question as a unit-cost economics test. On the AI side, cost per task decomposes into a variable inference cost dominated by electricity and the supply chain that produces compute, an amortised integration cost recoverable across volume, and a residual supervision cost that regulatory and professional regimes will not let fall to zero. Once these three lines are loaded, the substitution discourse turns out to operate as a feedback structure: capability gain stresses the supporting systems faster than it improves the calculus on which substitution depends. The variables outside the AI debate, energy markets, integration economics, professional regulation, capital-market patience, and the geopolitics of compute supply chains decide whether substitution clears in any given place at any given time. Substitution arrives selectively, configurationally, and at a more modest scale than the dominant discourse implies.

Keywords Artificial intelligence and labour · Capability determinism · Unit-cost economics · Inference cost · Technological determinism · Labour substitution

1 Introduction

The contemporary debate on artificial intelligence and labour has settled into a recognisable pattern across consultancies, multilateral institutions, and national policy frameworks. A model capability is demonstrated, an inference about workplace substitution is drawn from that demonstration, and the conversation moves quickly to the distributional and social-policy questions that the inferred substitution is taken to produce (Hatzius et al. 2023; Chui et al. 2023; Cazzaniga et al. 2024; OECD 2023; HM Government 2025; European Commission 2024). What is striking is that the underlying structure of the inference, from capability to substitution to distribution, has acquired the status of an unargued default rather than an argued position (Cole et al. 2022; Verdegem 2022; Willems et al. 2025).

This paper takes that default as its object of analysis. The subject matter is the regulatory and policy discourse on AI labour displacement; the conceptual framework is the older

sociology and history of technology literature on technological determinism (Heilbroner 1967; Winner 1980; Joerges 1999; Wyatt 2008), brought into productive contact with the operational accounting tradition on fully loaded cost comparison (Drury 2018; Horngren et al. 2014). The core argument advanced here is that the dominant framing of AI labour displacement is best understood as a form of *capability determinism* operating in a soft or normative register, and that an alternative analytic frame, the *unit-cost economics* of substitution, relocates the question to a comparison of fully loaded cost per task in which the AI side decomposes into three lines: a variable inference cost dominated by electricity and compute supply chains; an amortised integration cost recoverable across volume; and a residual supervision cost that regulatory and professional regimes will not let fall to zero.

The contribution to the existing regulatory discourse follows from these moves taken together. The contemporary conversation is pivoted towards the management of substitution consequences (Cazzaniga et al. 2024; OECD 2023; Susskind 2020) on the assumption that substitution will clear at the rate the capability rhetoric implies. The unit-cost economics frame relocates the analysis to the prior question

✉ Will Mbioh
w.r.mbioh@kent.ac.uk

¹ University of Kent, Canterbury, UK

of whether and where substitution will actually clear, and brings into the labour debate domains that the dominant framing has held outside it: energy regulation (IEA 2025, 2026), supply chain governance (Miller 2022; Khan 2021), professional regulation (Pasquale 2020; Susskind and Susskind 2015), and integration economics (Brynjolfsson et al. 2021; Davenport and Mittal 2023). The paper proceeds as follows. Section 2 sets out the dominant frame and reconstructs what labour substitution actually means as a category. Section 3 develops the unit-cost economics critique through the variable inference cost line. Section 4 takes the analysis up the stack to integration and supervision. Section 5 connects the analysis to the older technological determinism debate and concludes.

2 The dominant AI labour-substitution discourse

Two preliminary distinctions need to be drawn before the critique can proceed. The first concerns what *labour substitution* actually means as a category. The second concerns what the dominant discourse around AI is saying about it. The two are related but separable, and a great deal of confusion in the current debate follows from running them together.

2.1 What labour substitution means

Labour substitution is, fundamentally, an accounting question that becomes a political question. The accounting question is whether a firm, faced with a task currently performed by a human being, can replace that human being with something else, usually a machine, and produce the same output at the same or better quality for less money per unit of output. If the answer is yes, the firm has an economic incentive to substitute. If the answer is no, it does not.

The political question follows from the accounting question. Once enough firms substitute in enough places, the labour market for that category of work shrinks. People who used to do that work for a living find that there is less of it to do, that the wage for it has fallen, or that the work has disappeared altogether. The distributional consequences of that shrinkage become a political problem, because somebody has to decide what happens to those people: whether they are retrained, redistributed to, given basic income, allowed to fall through, or absorbed somewhere else in the labour market (Susskind 2020; Acemoglu and Johnson 2023).

The classical economic theory of substitution treats this as a story about *factor prices*. There are two factors of production in the simplest version, labour and capital, and the firm picks the mix that produces output most cheaply. When the price of capital falls relative to the price of labour, or

when capital becomes more productive per unit of cost, the firm substitutes capital for labour. Historically, this is the story of the steam loom replacing the handloom weaver, the assembly line replacing the artisan, the spreadsheet replacing the bookkeeper. The technology arrives, the unit cost of doing the work with the machine falls below the unit cost of doing it with the person, and the substitution occurs. The displaced workers find themselves needing either to acquire skills the new technology cannot replicate, or to move into sectors the technology has not reached.

The contemporary refinement of this story, associated particularly with task-based models of automation in labour economics, is more granular (Acemoglu and Restrepo 2022; Frey and Osborne 2017). It says the unit of substitution is not the *job* but the *task*. A job is a bundle of tasks, and a new technology typically substitutes for some tasks within the bundle whilst complementing or leaving alone others. So a given technology might displace twenty per cent of a paralegal's tasks without displacing the paralegal, and over time the bundle of tasks that defines the paralegal's job shifts towards the things the technology cannot yet do. The labour market consequences depend on how many tasks across how many jobs get substituted, and on how easily workers can move into the residual non-substitutable tasks.

That is the general theory. Substitution is firm-level cost calculation, summed up across enough firms and tasks to reshape the labour market, with distributional consequences that become political.

2.2 The capability discourse of AI

On top of that general theory, the specific way the contemporary discourse talks about AI must be layered in. The capability discourse is the way the AI-and-labour conversation has come to be framed in policy documents, consultancy reports, multilateral agency papers, and the headline coverage that filters all of these into the press.

The discourse begins from a *demonstration of model capability*. A new AI model is released, it can do something it could not do before, it scores higher on some benchmark, it passes a professional examination, it produces a piece of writing or code or analysis that compares favourably to what a human would produce (Eloundou et al. 2024; Noy and Zhang 2023). That demonstration is the lead. It is what the press conference is about, what the headline reports, what the executive summary leads with.

From the demonstration of capability, the discourse runs to an *inference about substitution*. If the model can do the task, the inference goes, then the model can replace the human who currently does the task. This step is where most of the work happens, and it is also where most of the assumptions get smuggled in. The discourse rarely makes the inference explicit, because making it explicit would

require defending it. Rather, it slides from “the model can do X” to “the model will do X at scale” without pausing in between. The Goldman Sachs projection that two-thirds of US occupations are exposed to some degree of AI automation (Hatzius et al. 2023), the McKinsey estimate of trillions of dollars of productivity uplift (Chui et al. 2023), the IMF projection of forty per cent of global employment exposed (Cazzaniga et al. 2024), the OECD’s framings of the future of work (OECD 2023; Georgieff and Hyee 2022), and the AI occupational exposure scores that underwrite all of them (Felten et al. 2021; Eloundou et al. 2024) are all instances of this slide. They take a measure of what models can do, multiply it by an exposure score, and report the result as if it were a forecast of what will happen.

From the inference about substitution, the discourse runs to a *distributional politics*. The conversation moves quickly to what should be done about the people who are about to be displaced. The instruments under discussion are retraining schemes, lifelong learning agendas, redistribution mechanisms, social-insurance reform, basic-income pilots, and tax-base reform for an economy that will have less labour income to tax. Susskind’s *A World Without Work* (Susskind 2020), Acemoglu and Johnson’s *Power and Progress* (Acemoglu and Johnson 2023), and the policy frameworks of the major jurisdictions (HM Government 2025; European Commission 2024) all do this. The substantive content is about managing the consequences of an event whose occurrence is treated as settled.

That is the structure. Capability is demonstrated, substitution is inferred from the capability, distributional politics follows from the inferred substitution. Three steps, and almost the entire conversation lives in the third one.

2.3 What the discourse skips

What can be observed, when labour substitution as a general theory is placed next to the capability discourse, is that the discourse skips the step the theory says is central. The theory says substitution is a firm-level cost calculation question: does the machine do the task more cheaply per unit, once every cost of running the machine has been counted? The discourse does not run that calculation. It treats the demonstration of capability *as if* it answered the cost question, and proceeds straight to the distributional consequences as if substitution at scale were already happening.

The cognitive form the discourse takes is what makes this slide possible. It treats AI as a *mind*, a disembodied intelligence whose capabilities can be compared to a human’s (Bender et al. 2021; Crawford 2021; Suchman 2007). It does not treat AI as a *system*, a stack of physical infrastructure that has to be powered, cooled, supplied with chips and water and electricity, integrated into a firm’s existing processes, and supervised by people who carry the regulatory

and professional liability for what it produces (Mosco 2014; Brodie 2024; Hogan 2015). Once AI is framed as a mind, the question of whether substitution clears looks like a question about whether the mind is good enough. Once AI is framed as a system, the question becomes whether the whole stack comes in below the cost of the human across enough deployments to matter. Those are different questions. The capability discourse asks the first one. The critique developed in this paper insists that the second one is what actually decides the outcome.

3 The variable inference cost: why electricity is the binding constraint

The argument now turns to one specific cost line, the cost of the electricity that powers the data centre, which ends up doing more work in deciding whether substitution actually clears than the capability discourse acknowledges. Three preliminary clarifications are required: what an inference actually is at the physical level, how a data centre produces inferences, and why electricity ends up being the binding constraint on the price of an inference and therefore on whether the substitution calculus comes out on the side of the machine or the human.

3.1 What an inference is, physically

When an AI model is asked to do something, draft a contract clause, summarise a document, write a piece of code, the technical name for what the model does in response is an *inference*. Inference is the running of the trained model. It is distinct from training, which is the one-off process by which the model is built in the first place. Training happens once per model, costs hundreds of millions of pounds, and produces a set of fixed numerical weights (Patterson et al. 2021; Luccioni et al. 2023). Inference happens every time the model is used after that, and it is what generates the output the user sees.

At the physical level, an inference is a long sequence of calculation operations performed on a specialised chip called a graphics processing unit, or GPU. The model’s response to a prompt is built up one *token* at a time, where a token is roughly a fragment of a word, a punctuation mark, or a short word in its entirety. To produce each token, the chip has to multiply very large grids of numbers together, run the result through the model’s weights, and select the next token. A short response of a hundred words might involve a hundred and fifty tokens. A longer reasoning task might involve thousands. A so-called agentic workflow, where the model calls itself repeatedly to plan, act, and revise, might involve tens of thousands (Wei et al. 2022; Snell et al. 2024). The number of tokens is the unit of work. Every token costs

the same kind of calculation, performed on the same kind of chip, drawing the same kind of power.

This matters because the question of what it costs to run an AI task is, at root, the question of how many tokens the task requires and what each token costs to produce. The token count varies with the task. The cost per token is set by the physics of the chip and by the price of the electricity flowing into it.

3.2 What a data centre is, physically

A data centre is the building that houses the chips. At the scale required for frontier AI inference, this is not a small facility. The current generation of AI data centres are measured in megawatts, sometimes gigawatts, of electrical capacity. To put that in scale, a one-gigawatt data centre draws roughly the same power as a medium-sized city. Globally, data centre electricity consumption has been growing at around twelve per cent per year since 2017, more than four times faster than total electricity consumption (IEA 2025; Ferriani and Gazzani 2026). The International Energy Agency expects this to continue (IEA 2026).

Inside the building, the chips sit in racks. The racks generate heat, a great deal of heat, because every calculation operation a chip performs is fundamentally an electrical event that releases energy as warmth. Modern AI chips, particularly the Nvidia H100 and H200 series and their successors, run hot enough that air cooling is no longer adequate at scale. The newer facilities use liquid cooling, where coolant runs through pipes inside the rack to carry the heat away (Mytton 2021). The cooling system itself consumes electricity, and at the scale of a hyperscale facility, it consumes a substantial fraction of what the chips themselves consume. The standard measure, called power usage effectiveness, expresses this as a ratio. A perfectly efficient facility would have a ratio of one, where every watt drawn from the grid goes to computation. A typical good facility runs at around 1.3, meaning thirty per cent of the power drawn from the grid is consumed by cooling, lighting, power conversion, and the other ancillary systems that keep the chips running (IEA 2025).

When a chip performs an inference, the electricity flowing into the chip is producing an numeral answer and producing heat. The heat then has to be removed, which requires more electricity. In some regions, the cooling also consumes water, sometimes very large quantities of it (Hogan 2015; Mytton 2021; Bresnihan and Brodie 2021). The chips themselves are bought from a small number of suppliers, principally Nvidia, which in turn depends on Taiwan Semiconductor Manufacturing Company for fabrication (Miller 2022). The chips contain rare materials whose supply depends on stable trade in helium, gallium, bromine, and a handful of

other inputs that come from politically concentrated sources (Khan 2021; World Economic Forum 2026).

Around all of this sits the rest of the stack. The building has to be constructed and maintained. The grid connection has to be negotiated with the local utility, which is increasingly the binding constraint on where new facilities can be built at all (IEA 2026; Ferriani and Gazzani 2026). The site has to be insured, which is a growing problem as insurers reprice the catastrophic loss exposures of facilities that can cost twenty billion dollars before any equipment is installed (Swiss Re Institute 2026). Staff have to be hired to operate the facility. None of these costs is negligible, but most of them are *fixed* costs, paid once or paid annually regardless of how many inferences the facility produces.

The cost per inference is what is obtained when the building, the chips, the cooling, the connection, the insurance, the staff, and the electricity are taken together and divided by the number of inferences the facility produces over its life. The electricity is the only one of those that scales directly with the work the facility does. Run more inferences, burn more electricity. Run fewer, burn less. The other costs are paid whether the facility is running flat out or sitting idle (Drury 2018; Horngren et al. 2014).

3.3 Why electricity ends up being the binding constraint

If the cost of producing a single inference is decomposed into its parts, two things become visible. First, the fixed costs only become small per inference if the facility is run at high utilisation, meaning if there is enough demand to justify keeping the chips busy most of the time. Second, the variable cost, the cost that scales with each inference, is dominated by electricity. The electricity bill of a hyperscale data centre is a number with nine or ten digits in it annually, and it is the largest single line item in operating the facility once it is built (IEA 2025; Patterson et al. 2021).

Two things follow from this. The first is that the long-run economic price of an inference is a function of the price of electricity. If electricity is cheap, inference can be cheap. If electricity is expensive, inference cannot be cheap, no matter how clever the model is or how efficiently the chips run. The chips can be improved, the cooling can be improved, the algorithms can be made more efficient, but at the bottom of the stack, there is a kilowatt-hour of electricity being converted into calculations and heat, and the price of that kilowatt-hour sets the floor on what the inference can cost (Boyd 2014; IEA 2025).

The second is that the price of electricity is not stable, and is moving in directions that make the floor higher rather than lower. In the United States, household electricity prices are now around thirty-two per cent above their 2019 levels, and the structural drivers behind that increase are not

transient (IEA 2026; Ferriani and Gazzani 2026). Data centre demand is itself one of those drivers. Because data centres are concentrated in particular regions, the local pressure on electricity prices in those regions is much sharper than the national average (Bresnihan and Brodie 2021; Ferriani and Gazzani 2026). The grid in these regions was built to serve the demand that was anticipated before the AI build-out, and adding several gigawatts of new load to it, in a few specific places, creates the kind of localised price pressure that the wholesale market then transmits into wider rates. On top of that base trend, geopolitical shocks to the underlying energy markets, such as those associated with the Iran war of 2026 and the disruption to oil and gas flows through the Strait of Hormuz, push the cost of the fuels that generate electricity upward, and the data centre's grid bill rises with them (Chatham House 2026; Foreign Policy 2026; World Economic Forum 2026).

3.4 How this then decides whether the substitution clears

This must now be placed back next to the substitution equation. A firm considering whether to replace some category of cognitive labour with AI is running, whether it knows it or not, a comparison between the cost per unit of human work and the cost per unit of machine work. The cost per unit of machine work depends on the price the firm is paying per inference, multiplied by the number of inferences each unit of work requires.

Today, the price per inference looks very low. The frontier providers are charging API prices that work out to a few pennies per task; API prices are the per-token rates published by the model providers, payable by any firm or developer that sends requests to the model through its application programming interface (API), the technical channel through which the model is accessed at scale rather than through a chat window. This makes the substitution look easy. But the price the providers are charging today is not the long-run economic price of producing the inference. It is a subsidised price, funded by capital raised on the strength of forecasts of future demand and future cost reduction. The Epoch AI analysis of OpenAI's GPT-5 bundle finds gross margins on inference of around thirty per cent and operating losses once everything else is included (Sevilla, Petrovic and Ho 2026). The frontier providers are losing money on inference. They can do this for as long as the capital markets are willing to fund the losses, on the bet that demand will eventually be large enough and costs will eventually be low enough that the unit economics will close (Birch and Cochrane 2022).

If the capital markets eventually demand an economic price, the price per token rises. If electricity prices rise, the floor under the price per token rises with them. If both happen simultaneously, the price per token rises sharply. And

several other things are happening at the same time that push the number of tokens per task in the wrong direction for the substitution. The newer reasoning models, which produce better answers on hard tasks, do so by burning more tokens per task, sometimes ten or twenty times more (Snell et al. 2024; Wei et al. 2022). Agentic workflows, where the model plans and acts in multiple steps, multiply token consumption again. Retrieval-augmented systems, which feed extra context into the model to improve its grounding, also multiply tokens. Hence, the very techniques that make AI better at the tasks the substitution is targeting are also making each instance of that task more expensive to run (Patterson et al. 2021; IEA 2026).

These factors must be stacked together. The price per token has a structural floor set by electricity. Electricity is rising. The number of tokens per task is rising too, because better answers are bought at the cost of more compute. The promotional price the firm is paying today is below the economic price and will eventually have to converge on it. And the economic price itself is being set by an input, the cost of electricity, that is determined in markets the firm has no control over: energy markets, grid policy, geopolitics, supply chains for industrial gases used in chip fabrication (Khan 2021; Miller 2022).

This is why the substitution question, when it is loaded up with the full cost of running the machine, is a question about electricity as much as it is a question about capability. The model can be capable enough to do the task. The model can get more capable still. But if the price of producing an inference rises faster than the price of paying a human to do the same work, then the substitution calculation does not clear, the firm keeps the human, and the labour displacement that the capability discourse forecasts at scale simply does not happen at scale.

The capability discourse cannot see this, because it does not look at the bottom of the stack. It treats AI as a mind whose performance is the relevant variable, and treats the infrastructure as a background condition that will sort itself out (Crawford 2021; Mosco 2014). The infrastructure is not a background condition. It is the binding constraint. And the price of the electricity flowing through it is the number that decides whether substitution actually happens or whether it remains, at scale, a forecast that never quite arrives.

4 Integration and supervision: why scale and accountability break the substitution maths

The analysis will now move up the stack into two further cost lines that the headline price of an API call hides. The first is what happens when an organisation tries to move from a few staff using AI as a personal productivity tool

to actually replacing categories of work with AI at scale, particularly with agentic systems. The second is what happens when the work is in a sector where somebody senior, expensive, and professionally accountable has to check what the AI has produced before it leaves the building. Both are cost lines that the current discourse mostly does not see, and both can knock a substitution decision out even when the headline cost per inference looks negligible.

4.1 The integration problem and why agents make it dramatically worse

When a single lawyer or analyst or coder uses ChatGPT or Claude on their own laptop to help with a specific task, the cost is the subscription, perhaps twenty pounds a month, plus the time the person spends prompting and checking the output. That is not substitution. That is a productivity tool. The lawyer is still doing the work. The AI is helping at the margin (Brynjolfsson et al. 2025; Noy and Zhang 2023). The economics of this are workable because the AI is essentially free relative to the lawyer's hourly rate, and the only question is whether the lawyer is faster with it than without.

Substitution is a different proposition. Substitution means the firm is no longer paying the lawyer to do the work. It means the AI is producing the work output that previously came from a person, with the firm taking responsibility for that output, and with the firm's processes restructured so the AI's work flows through the same channels the person's work used to flow through. Getting from "we have a chatbot" to "we have actually replaced the work" requires a great deal of building, and this is what people in the industry call *integration* (Davenport and Mittal 2023; Brynjolfsson et al. 2021).

Integration involves several things that are individually unglamorous and collectively expensive. The AI has to be connected to the firm's data, which means the firm's documents, case files, customer records, financial data, code repositories or whatever the relevant inputs are. That data has to be cleaned, indexed and made retrievable, because raw data sitting in a thousand folders across a company server is not in a form an AI can work with. The AI has to be wired into the firm's existing software systems, the case management system at a law firm, the practice management system at a hospital, the customer relationship management system at a sales operation. Permissions have to be set up so the AI can see what it is supposed to see and not see what it is not supposed to see. Audit trails have to be built so the firm can show, when something goes wrong, which inputs the AI saw and what it produced. The AI's outputs have to flow back into the firm's existing processes in formats that downstream people and systems can use. Staff have to be trained on the new workflow. Procedures have to be rewritten. Compliance

has to sign off. Insurance has to be reviewed (Davenport and Mittal 2023; Hacker 2023).

This work is done by some combination of internal engineers, external consultants, and the AI vendor's professional services team. For a serious enterprise deployment of any depth, the bills run into the multiple millions and the timelines run from 1 to 2 years before the system is reliably in production (Chui et al. 2023; Davenport and Mittal 2023). The numbers are not surprising. They are the kind of numbers any major enterprise software deployment generates; AI is not different from any other system's integration project in this respect; and the costs are amortised over the working life of the deployment (Brynjolfsson et al. 2021).

The *amortisation* point is the one that matters for the substitution calculation. If a firm spends three million pounds integrating an AI system, and uses it to do work that previously cost ten million pounds a year in salaries, then the integration cost adds about thirty pence per pound of work in the first year and falls to a much smaller number in subsequent years if the deployment is durable. The calculation clears. If the same firm spends three million pounds integrating a system to do work that previously cost five hundred thousand pounds a year, then the integration cost adds six pounds for every pound of work. The calculation does not clear. The firm would have to keep the deployment running for many years before the integration cost was absorbed. The decision to integrate at scale, therefore, depends on whether the firm has enough volume of the substituted task to spread the integration cost over (Drury 2018; Horngren et al. 2014).

This is one reason why predictions of mass white-collar substitution often miss. A great many firms do not have the volume of any particular task to justify the integration. They have a hundred different tasks done in small quantities by people who are flexible across all of them, and the cost of integrating an AI for any one of those tasks would not be recovered. The substitution that makes economic sense is concentrated in firms with very high volumes of a single repeatable task, which is why the early successes of enterprise AI cluster in customer support, content moderation and certain kinds of standardised analytical work (Brynjolfsson et al. 2025), and why deployment in the more variegated knowledge professions has been much patchier (Cole et al. 2022; Willems et al. 2025).

Agents must now be layered on top of this, because this is where the cost trajectory turns sharply against the substitution. An *agent* is an AI system that can plan, act and revise across many steps, calling itself or other tools repeatedly to accomplish a multi-stage task (Wei et al. 2022; Snell et al. 2024). Where a single inference might use a few hundred tokens, an agent doing a complex task can use tens of thousands or hundreds of thousands. Each step of the agent's reasoning is itself an inference. Each tool call the agent makes generates more inferences when the result comes back. The

agent often re-reads the entire history of the task to decide what to do next, which means the same context gets fed through the model again and again, with each pass costing tokens.

The token consumption of an agentic system relative to a single chat call is not double or triple. It is one or two orders of magnitude higher per task. So if a basic chat inference costs five pence per task, the agent doing the same task end to end might cost five pounds, or in some cases fifty (Snell et al. 2024; Patterson et al. 2021). The headline pricing of the API was built on the assumption of conversational use, where the user asks a question and the model answers. Once the user is no longer in the loop, and an agent is running for hours autonomously, the cost per task can quickly become comparable to or higher than the cost of a junior employee doing the work in the same time.

There is a second compounding effect. Agents are unreliable in ways that single inferences are not. When an agent runs for fifty steps, the probability of an error compounds across those steps. A small error in step seven gets carried into the inputs for step eight, and into the inputs for step nine, and the agent can spend large amounts of token budget pursuing a wrong path before catching itself or before a human intervenes (Bender et al. 2021). Hence, the average cost per successful task is higher than the cost per task in the cases where everything goes right. Empirically, organisations attempting agentic deployments find the *variance* in cost per task is substantial, which means they have to budget for the bad cases as well as the good ones. And the bad cases burn token budget without producing usable output, which means they sit on the cost side of the calculation without anything on the benefit side.

The picture, when an organisation tries to move from individual productivity use to substitution at scale with agents, is that two things happen simultaneously. The integration cost has to be recovered against the volume of the deployed task, which often does not work for the most variegated parts of the knowledge economy. And the per-task cost of running the agent is much higher than the per-task cost of the underlying inference, sometimes orders of magnitude higher, with significant variance and significant token burn on tasks the agent fails to complete. Both of these together mean that the substitution that looked easy from the headline API price gets much harder once the firm tries to do it for real.

4.2 The supervision problem as a significant factor

The third cost line is the one the discussion of capability misses entirely. This is the cost of the human who has to check what the AI has produced before it leaves the firm. In some sectors, this cost is small, because the consequence of an error is low and the output is not subject to professional

or regulatory accountability. In many sectors, including the ones that the substitution discourse points at most often as targets, this cost is large, sometimes large enough on its own to defeat the substitution calculation (Pasquale 2020; Susskind and Susskind 2015).

Consider law. A junior associate's draft of a contract clause has historically gone to a partner for review before it goes to the client. The partner reviews it because the partner's name is on the legal advice, the partner's professional registration is on the line if the advice is wrong, the firm's professional indemnity insurance prices the cover on the assumption that partners review junior work, and the client is paying partner rates partly to get partner judgement on what the juniors produce (Susskind and Susskind 2015; Hacker 2023). Suppose now that the AI replaces the junior. The draft still has to go to the partner, because nothing in the supervisory regime has changed. The regulator still requires partner sign-off on advice that goes out under the firm's name. The insurer still requires the same review structure. The client still expects the same standard of professional care. The cost the substitution was supposed to eliminate, the price of senior expert time, has not been eliminated. It has been moved from drafting and supervising the junior to reviewing and editing the AI's output. And in many cases, the partner spends as long reviewing AI work as they would have spent reviewing junior work, because they cannot trust the AI's judgement without checking, and because spotting where an AI has subtly hallucinated a citation or misread a clause requires the same expertise that producing the work in the first place would have required (Pasquale 2020; Bender et al. 2021).

Medicine works the same way, more sharply. If an AI system suggests a diagnosis or a treatment plan, somebody with a medical licence has to take responsibility for the decision before it affects the patient (Pasquale 2020). The licence holder is the supervisory cost, and the supervisory cost does not disappear because the suggestion came from a machine rather than from a registrar.

Audit and finance work similarly. The signing partner on an audit takes personal responsibility for the audit opinion. AI tools can speed up the work that goes into forming the opinion, but the opinion itself has to be signed by a human who is professionally and legally accountable, and that human has to have done enough of their own work on the file to defend the signature (Susskind and Susskind 2015; Hacker 2023). The supervisory cost is the cost of that human's time.

Coding is more interesting because people often assume it is exempt. It is not, but the way the supervisory cost shows up there is different. In safety-critical software, in financial systems, in code that runs in regulated industries, and in any code that will go into production at a serious organisation, the human who reviews the code retains professional

responsibility for what it does once it is deployed. Code review is the supervisory mechanism. Senior engineers reviewing AI-generated code spend substantial time understanding what the AI has done, checking for subtle bugs, verifying the code does what it was supposed to do, and ensuring it does not introduce security vulnerabilities (Brynjolfsson et al. 2025). Empirically, AI coding tools speed up some tasks dramatically and create review burdens that partially offset the speed gains. The net productivity benefit is real but is much smaller than the headline numbers suggest, particularly for code that has to meet a high reliability bar. In contexts where the code is throwaway, a quick prototype, an internal script, a personal tool, the supervisory cost falls towards zero and the productivity gains are larger. But this is precisely the work that was cheapest to do with humans in the first place, so the substitution there does not displace much labour cost (Acemoglu and Restrepo 2022).

The general pattern is that the supervisory cost is highest exactly in the sectors where the labour cost the substitution targets is also highest. The professions whose senior practitioners earn the most are also the professions whose regulatory and liability regimes require the most senior review (Pasquale 2020; Susskind and Susskind 2015; Hacker 2023). Hence, the substitution that would have been most valuable in absolute terms, replacing high-cost cognitive labour, is the substitution where the supervisory cost is least eliminable. And the substitution that would have been least valuable, replacing low-cost work where supervision is light, is precisely where the substitution is technically easy but financially uninteresting.

The three cost lines stacked together produce a clear picture. The variable inference cost is dominated by electricity and is moving up. The integration cost is large, has to be amortised across volume, and is sharpest for agentic deployments where token consumption explodes. The supervision cost is largest exactly where the substituted labour is most expensive. The headline API price, which is what the capability discourse implicitly multiplies by the number of substitutable tasks to get its forecasts, is the smallest of the three cost lines and the only one that is currently subsidised. When the firm tries to do the substitution at the scale and in the sectors the discourse points at, all three lines come up at once, and very often at least one of them is enough to keep the human in the job.

5 Capability determinism and what comes instead

All of this challenges the form of technological determinism that is implied by much of the substitution discourse. To understand why, a working sense of what technological determinism is as a position is required. The phrase comes

out of a long argument in the history and sociology of technology, but at its core it names a way of thinking about the relationship between technology and society that gets the direction of causation wrong, or at least gets it more wrong than right (MacKenzie and Wajcman 1999; Wajcman 2010). The deterministic position holds that technologies, once invented, more or less make the world around them. The arrow runs from the technology to the social arrangements. The technology is the active force, society is what gets reshaped by it. The classical formulation of this question, posed by Robert Heilbroner in 1967 in an essay entitled “Do machines make history?” (Heilbroner 1967), asks whether the development of technology follows an autonomous logic that then produces the social structures around it, or whether the causal arrow runs the other way, with social and political and economic conditions shaping which technologies get developed and how they get deployed.

There are stronger and weaker versions of this position. The strongest version, which Langdon Winner explored in his 1980 essay “Do artifacts have politics?” (Winner 1980), holds that technologies have politics built into them, in the sense that the design of a piece of technology can embed political commitments that then get realised wherever the technology is used. A bridge built deliberately too low for buses to pass under, in Winner’s much-debated example, has the politics of racial exclusion built into its concrete because the people who could not afford cars could not get to the beach the bridge led to. A weaker version, which the sociologist Sally Wyatt articulated in a 2008 chapter entitled “Technological determinism is dead, long live technological determinism” (Wyatt 2008), concedes that hard determinism cannot survive contact with the historical record, because the record obviously contains too much resistance, regulation, refusal and redirection of technologies for the claim that technologies simply produce their social outcomes to be true. But Wyatt argued that determinism keeps coming back in a softer or normative form, in which the *rhetoric* of technological inevitability does the determining work even though the technology itself does not (Wyatt 2008; Joerges 1999). The rhetoric structures how options are perceived. It forecloses alternatives before they reach the table. It makes resistance look like denial of an obvious future rather than a legitimate political choice (Mueller 2021; Cole et al. 2022). The Heilbroner question, in Wyatt’s hands, becomes: even if machines do not make history, do statements about what machines will do make history, by shaping the political and economic decisions that determine what gets built and what gets done with it?

The substitution discourse around AI sits inside this softer version. It rarely argues, in its strongest form, that AI capability *causes* labour displacement. Rather, it argues that AI capability *will produce* labour displacement, and the productive force is treated as something close to automatic. The

Goldman Sachs projection (Hatzius et al. 2023), the McKinsey estimate (Chui et al. 2023), the IMF discussion note (Cazzaniga et al. 2024), the OECD framing (OECD 2023), the national policy frameworks (HM Government 2025; European Commission 2024) all share the same structure: a measure of what AI can do is multiplied by an exposure score, and the result is reported as a forecast of what will happen. The forecast is then taken as a premise on which the policy conversation builds. Retraining schemes are designed for the displaced. Tax-base reform is debated for the post-labour economy. Basic-income pilots are discussed for the world the substitution will produce (Susskind 2020; Acemoglu and Johnson 2023). The capability claim, having entered the discourse as a description of what models can do in a benchmark, exits it as a description of what the labour market will look like in 5 or 10 years.

This is what makes the discourse deterministic in *the soft sense*. It is not that anybody is arguing the technology has a will of its own. Rather, it is that the variables which would actually decide whether the substitution happens, the ones traced in the previous sections, are treated as background conditions that the system will route around rather than as binding constraints on the outcome. Energy prices, integration costs, supervisory regimes, supply chains, professional regulation, the cost of capital that is currently subsidising the inference price, the geography of grid capacity, the politics of the regions where the chips and the gases and the choke-points sit, all of these get bracketed (Crawford 2021; Brodie 2024; Verdegem 2022). The capability claim moves to the front of the argument, the substitution claim is inferred from it, and the distributional claim is built on top of the substitution claim. The variables that would determine whether the inference is sound have been moved offstage.

Once the cost stack is understood, the deterministic move becomes much harder to sustain. The price of electricity is not a background condition; it is the largest single line in the variable cost of producing an inference, and it is set in markets the AI sector does not control (IEA 2025, 2026; Boyd 2014). The integration cost is not a background condition; it is several million pounds and a year or two of work that has to be amortised against the volume of the deployed task, and many of the firms whose substitution decisions the forecasts are aggregating do not have the volume to support the amortisation (Davenport and Mittal 2023; Brynjolfsson et al. 2021). The supervisory cost is not a background condition; it is the price of senior expert time in regulated professions, and the regulatory regimes that require it are not adjusting to capability improvements at the rate the capability discourse assumes (Pasquale 2020; Hacker 2023). Each of these is a variable that can move independently. Each of them can move in directions that prevent the substitution from clearing. And several of them are moving in those directions simultaneously: electricity prices rising (IEA 2026); capital

markets approaching the point at which they will demand an economic price for inference (Sevilla, Petrovic and Ho 2026; Birch and Cochrane 2022); agentic systems multiplying token consumption per task (Snell et al. 2024); professional regulation tightening rather than loosening around AI outputs (Hacker 2023).

The question of whether AI labour substitution happens at the scale the discourse implies is therefore not a question whose answer follows from a measurement of capability. It is a question whose answer depends on the trajectory of energy markets, the resilience of supply chains for chips and industrial gases, the volume economics of integration in the specific firms making the decision, the supervisory burden the regulator imposes, the price the capital markets are willing to keep paying for losses on the supply side, and the geographies of grid capacity that determine where new inference can be built at all. None of these is determined by capability alone. Some of them are moving against the substitution rather than with it.

That is, the capability discourse cannot accommodate within its own framing. In that, the discourse imagines a single curve along which all the variables travel together: capability improves, costs fall, integration matures, substitution clears. Each variable is treated as downstream of capability. The cost stack suggests the opposite arrangement. Several of these variables are responding to AI's own success by moving against the substitution. Reasoning models burn ten or twenty times more tokens per task than chat models did; agentic workflows multiply this further; the very advances that make substitution attractive in principle raise its cost in practice. Electricity prices rise partly because AI demand tightens the regional grids it depends on. Professional regulation tightens partly because AI capability invites scrutiny. Capital markets, having subsidised inference on the bet that scale would close the unit economics, must at some point confront the prospect that scale alone will not close it. The substantive claim is that AI's progress generates headwinds in the systems on which its own substitution logic depends. The capability story is single-loop: capability rises, substitution clears. The reality is a feedback structure in which capability gain stresses the supporting systems faster than it improves the underlying function.

In other words, two rates are running simultaneously inside the system. The first is the rate at which capability improvement makes inference cheaper per unit of quality, through better algorithms, more efficient chips, and more compact model architectures. The second is the rate at which capability improvement increases strain on the systems that produce inference: more tokens per task as reasoning and agency scale, more electricity demand as the buildout multiplies, more regulatory scrutiny as the stakes of AI errors rise in visibility, more capital-market exposure as the bet on scale-as-payoff lengthens. The capability discourse silently

assumes the first rate dominates the second. The empirical situation is that the second rate is dominating the first.

The gap between two rates is not a static lag that will eventually close. If the stress rate exceeds the rate at which the calculus improves, the gap widens as time passes. The substitution that the discourse projects as 5 or 10 years away sits on a trajectory whose horizon recedes the more capability advances. What appears in the forecast as an event approaching at a calculable rate is in reality an event whose conditions of possibility are moving further from satisfaction.

The structural reason the rates run in opposite directions lies in the fact that the system that improves the calculus and the system that produces the strain are not the same system. The calculus improves through laboratory innovation, where iteration cycles are short and the inputs (researcher hours, compute, training data) can be scaled at will. The strain accumulates through infrastructure, regulation and capital, where the inputs (planning consents, regulatory consensus, investor patience) cannot be scaled by command. The first system has discretion over its own pace; the second does not. The faster the first one runs, the more it loads the second one with demands the second cannot deliver. Capability is endogenous to AI; the supporting systems are exogenous. Endogenous improvement loaded against exogenous supply produces divergence.

The forecast treats the two rates as commensurable, computes a date, and dresses the date in the language of inevitability. The system underneath the forecast does not behave that way. It behaves like a system whose fastest part produces more demand than its slowest parts can absorb.

5.1 Conclusion

What this means looking forward is that the capability discourse is almost certainly overstating the scale and underspecifying the heterogeneity of what will happen. The two-thirds of US or UK occupations displaced in the timeframe the headlines imply will likely not materialise, because the cost stack does not support displacement at that scale at the price points the displacement requires. Substitution will clear in some places and not in others, and the pattern of where it clears will be governed less by which tasks the models are best at and more by which firms have the volume, the cheap power, the light supervisory regime, and the integration appetite to make the calculation work.

What will this look like in practice? Substitution will concentrate where the cost stack favours it: high-volume repeatable tasks, low supervisory burden, access to cheap electricity, mature integration paths. Customer support at scale, content moderation, certain kinds of standardised analytical work, marketing copy, basic translation, transcription (Brynjolfsson et al. 2025; Gray and Suri 2019; Tubaro

et al. 2020). Substitution will be patchy and partial in the variegated knowledge professions, where supervisory cost is high and volume is fragmented. Law, medicine, audit, regulated finance, senior engineering (Pasquale 2020; Susskind and Susskind 2015; Hacker 2023). Substitution will track electricity geographies, with the firms that can lock in cheap power gaining a substitution advantage and the firms in expensive grid regions finding the calculation harder to clear (IEA 2025, 2026; Ferriani and Gazzani 2026). And substitution will track regulatory geographies, with jurisdictions that adjust their professional regimes to allow lighter supervision moving faster than those that do not (Hacker 2023; Pasquale 2020).

What the capability discourse misses, in other words, is that the determinants of AI labour displacement are mostly outside the AI debate as currently constituted. They are in energy policy, grid regulation, supply chain governance, professional regulation, the cost of capital, and the political stability of the regions in which the chips and the gases and the chokepoints sit (Boyd 2014; Khan 2021; Miller 2022; Crawford 2021). The substitution that does happen will be shaped by those determinants, not by capability alone. And the policy conversation that has been built on the assumption of capability-driven substitution at scale will need to be rebuilt around the actual variables that will decide the outcome.

Author contribution W.R.M. is the sole author of the manuscript. W.R.M. was responsible for the conceptualisation, methodology, formal analysis, writing of the original draft, and review and editing of the final manuscript. W.R.M. read and approved the final version submitted for peer review.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acemoglu D, Johnson S (2023) Power and progress: our thousand-year struggle over technology and prosperity. PublicAffairs, New York
- Acemoglu D, Restrepo P (2022) Tasks, automation, and the rise in U.S. wage inequality. *Econometrica* 90:1973–2016
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 610–623
- Birch K, Cochrane DT (2022) Big tech: four emerging forms of digital rentiership. *Sci Cult* 31:44–58
- Boyd W (2014) Public utility and the low-carbon future. *UCLA Law Rev* 61:1614–1710
- Bresnihan P, Brodie P (2021) New extractive frontiers in Ireland and the moebius strip of wind/data. *Environ Plan E Nature Space* 4:1645–1664
- Brodie P (2024) Smarter, greener extractivism: digital infrastructures and the harnessing of new resources. *Info Commun Soc* 28:1061–1080
- Brynjolfsson E, Rock D, Syverson C (2021) The productivity J-curve: how intangibles complement general purpose technologies. *Am Econ J Macroeconomics* 13:333–372
- Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *Q J Econ* 140:889–942
- Cazzaniga M, Jaumotte F, Li L, Melina G, Panton AJ, Pizzinelli C, Rockall EJ, Mendes Tavares M (2024) Gen-AI: Artificial Intelligence and the future of work. IMF Staff Discussion Note SDN/2024/001. International Monetary Fund, Washington, DC
- Chatham House (2026) US energy prices were set to rise long before the Iran war. *Chatham House Expert Comment*, 11 March
- Chui M, Hazan E, Roberts R, Singla A, Smaje K, Sukharevsky A, Yee L, Zimmel R (2023) The economic potential of generative AI: the next productivity frontier. McKinsey Global Institute, New York
- Cole M, Cant C, Ustek Spilda F, Graham M (2022) Politics by automatic means? A critique of artificial intelligence ethics at work. *Front Artif Intell* 5:869114
- Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven
- Davenport TH, Mittal N (2023) All-in on AI: how smart companies win big with artificial intelligence. Harvard Business Review Press, Boston
- Drury C (2018) Management and cost accounting, 10th edn. Cengage Learning, Andover
- Eloundou T, Manning S, Mishkin P, Rock D (2024) GPTs are GPTs: labor market impact potential of LLMs. *Science* 384:1306–1308
- European Commission (2024) AI and the future of work. European Commission, Brussels
- Felten EW, Raj M, Seamans R (2021) Occupational, industry, and geographic exposure to artificial intelligence. *Strateg Manag J* 42:2195–2217
- Ferriani F, Gazzani A (2026) Powering the digital economy: The global expansion of data centres and its energy implications. *VoxEU Column*, 28 March. Centre for Economic Policy Research, London
- Foreign Policy (2026) Iran war, Hormuz closure hit helium and semiconductor supply chain. Foreign Policy
- Frey CB, Osborne MA (2017) The future of employment: how susceptible are jobs to computerisation? *Technol Forecast Soc Change* 114:254–280
- Georgieff A, Hye R (2022) Artificial intelligence and employment: new cross-country evidence. *Front Artif Intell* 5:832736
- Gray ML, Suri S (2019) Ghost work: how to stop Silicon Valley from building a new global underclass. Houghton Mifflin Harcourt, Boston
- Hacker P (2023) The European AI liability directives: critique of a half-hearted approach and lessons for the future. *Comput Law Secur Rev* 51:105871
- Hatzius J, Briggs J, Kodnani D, Pierdomenico G (2023) *The Potentially Large Effects of Artificial Intelligence on Economic Growth*. Goldman Sachs Global Economics Analyst, 26 March. Goldman Sachs, New York
- Heilbroner RL (1967) Do machines make history? *Technol Cult* 8:335–345
- HM Government (2025) AI opportunities action plan. HMSO, London
- Hogan M (2015) Data flows and water woes: the Utah Data Center. *Big Data Soc* 2:1–12
- Horngren CT, Datar SM, Rajan MV (2014) Cost accounting: a managerial emphasis, 15th edn. Pearson, Upper Saddle River
- IEA (2025) Energy and AI. International Energy Agency, Paris
- IEA (2026) Key questions on energy and AI. International Energy Agency, Paris
- Joerges B (1999) Do politics have artefacts? *Soc Stud Sci* 29:411–431
- Khan SM (2021) Securing semiconductor supply chains. Center for Security and Emerging Technology, Washington
- Luccioni AS, Viguier S, Ligozat A-L (2023) Estimating the carbon footprint of BLOOM, a 176B parameter language model. *J Mach Learn Res* 24:1–15
- MacKenzie D, Wajcman J (eds) (1999) The social shaping of technology, 2nd edn. Open University Press, Buckingham
- Miller C (2022) Chip war: the fight for the world's most critical technology. Scribner, New York
- Mosco V (2014) To the cloud: big data in a turbulent world. Paradigm, Boulder
- Mueller G (2021) Breaking things at work: the Luddites are right about why you hate your job. Verso, London
- Mytton D (2021) Data centre water consumption. *Npj Clean Water* 4:11
- Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381:187–192
- OECD (2023) OECD employment outlook 2023: artificial intelligence and the labour market. OECD Publishing, Paris
- Pasquale F (2020) New laws of robotics: defending human expertise in the age of AI. Harvard University Press, Cambridge
- Patterson D, Gonzalez J, Le Q, Liang C, Munguia L-M, Rothchild D, So D, Texier M, Dean J (2021) Carbon emissions and large neural network training. [arXiv:2104.10350](https://arxiv.org/abs/2104.10350)
- Sevilla J, Petrovic H, Ho A (2026) Can AI companies become profitable? *Epoch AI Gradient Updates*, 28 January (revised 6 March 2026). <https://epoch.ai/gradient-updates/can-ai-companies-become-profitable>
- Snell C, Lee J, Xu K, Kumar A (2024) Scaling LLM test-time compute optimally can be more effective than scaling model parameters. [arXiv:2408.03314](https://arxiv.org/abs/2408.03314)
- Suchman L (2007) Human-machine reconfigurations: plans and situated actions. Cambridge University Press, Cambridge
- Susskind D (2020) A world without work: technology, automation and how we should respond. Allen Lane, London
- Susskind R, Susskind D (2015) The future of the professions: how technology will transform the work of human experts. Oxford University Press, Oxford
- Swiss Re Institute (2026) Sigma insights 07/2026: Insuring AI - data centre value accumulation risks. Swiss Re Institute, Zurich
- Tubaro P, Casilli AA, Coville M (2020) The trainer, the verifier, the imitator: three ways in which human platform workers support artificial intelligence. *Big Data Soc* 7:1–12

- Verdegem P (2022) Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. *AI Soc* 39:727–737
- Wajcman J (2010) Feminist theories of technology. *Camb J Econ* 34:143–152
- Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst* 35:24824–24837
- Willems T, Hotan DJ, Tang JC, Norhashim NHB, Poon KW, Goh ZAG, Vinod R (2025) Assessing employment and labour issues implicated by using AI. In: Yaghmaei E et al (eds) *Global perspectives on AI impact assessment*. Oxford University Press, Oxford
- Winner L (1980) Do artifacts have politics? *Daedalus* 109:121–136
- World Economic Forum (2026) The strait of Hormuz crisis: rewriting the future of AI. World Economic Forum
- Wyatt S (2008) Technological determinism is dead; long live technological determinism. In: Hackett EJ, Amsterdamska O, Lynch M, Wajcman J (eds) *The handbook of science and technology studies*, 3rd edn. MIT Press, Cambridge, pp 165–180

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.