AI Is Not a Market and Treating It Like One Loses Cases

"A handful of firms control AI."

The sentence recurs in complaints, in submissions, and in conference keynotes: a handful of firms control AI. It is not false. The same few names really do recur, up and down the technology, with a frequency that should trouble anyone who cares about competitive markets.

But the sentence states the problem in a form no competition authority can act on, and it is worth spelling out why. Competition law does not ask whether a company is large, or powerful, or important in the abstract. It asks whether a company has power in a *market*, meaning a defined group of products that customers treat as substitutes for one another. Almost everything else follows from that definition. It determines whether a firm counts as dominant, whether a merger is reviewed, and whether conduct counts as an abuse or merely as vigorous competition. Define the market wrongly and the case fails, however troubling the underlying facts may be.

"AI" is not a market. It is a stack of four markets resting on one another, and a claim that is true of one layer is routinely false of the layer above it.

A short recap for new readers: This series argues that AI dissolves competition law's silent assumptions rather than breaking its rules. The previous article set out the map, and this one starts on the terrain.

The Four Layers

At the bottom of the stack sits compute. This is the physical foundation of the whole technology: the specialised chips that train and run AI models, the data centres that house those chips, and the cloud providers that rent access to both. Nothing above this layer exists without it.

Above compute sit the foundation models. These are the large, general-purpose systems that most people mean when they say "AI". They are expensive to build, they are trained on vast quantities of data, and they supply the raw capability that everything above them draws on.

Above the models is the agent layer. This is where general capability becomes something a person actually uses: the assistants and autonomous agents that answer questions and carry out tasks, together with the defaults that decide which system responds when a user simply asks. This is the layer the user touches.

At the top are the downstream applications. These are the ordinary software products that happen to use AI for some particular purpose, which is to say most of the software economy.

Those four descriptions are the map. Each layer has a different competitive story, and the differences are the point.

Compute is heavily concentrated, and it is likely to stay that way. The model layer looks concentrated but is contested in practice, because several laboratories (including the open source ones in China) sit at or near the frontier, the lead changes hands, and whether it ever settles is the most disputed question in the field. The agent layer is concentrating *conditionally*, which means the forces that entrenched search engines and web browsers, principally the power of being the pre-installed default, apply at least as strongly to assistants, but the consolidation has not hardened yet. Downstream is, for the most part, ordinary competition, with many firms, familiar churn, and nothing that requires special legal apparatus.

Where the Bottleneck Actually Sits

The rest of this article does two things with that map. It locates the one genuinely durable bottleneck in the stack, and it identifies which part of competition law attaches to each layer.

The bottleneck is not where the popular alarm points, and getting this right is what wins or loses cases. The alarm points at the chips, and the chip story is genuinely dramatic. On third-party estimates, since NVIDIA publishes no official figure of its own, a single firm supplies roughly three-quarters to four-fifths of the world's AI accelerators, which are the specialised processors that train and run these models.

But that dominance has two components, and they are moving in opposite directions.

The first component is the hardware itself, and its position is eroding. A second serious supplier has reached meaningful scale. The large cloud companies are designing their own chips, and those are growing several times faster than the general-purpose processors sold on the open market. There are reports this year that such custom chips may be sold to outside buyers, which would matter more than it first appears. A chip that anyone can buy separates the supply of computing power from the supply of cloud services, and that is new competition arriving *underneath* the cloud layer rather than reinforcement of it.

The strongest argument against that reading is a technology called NVLink Fusion, which allows cloud companies to plug their own custom chips into NVIDIA's surrounding hardware architecture. NVIDIA therefore stays embedded in the system even where its chip is not the one doing the main work. New entry beneath the cloud layer and continued dependence on the incumbent's connective technology are not mutually exclusive, and the honest position holds both at once.

The second component is not eroding at all, and this is the part that matters. It is the software ecosystem. NVIDIA's CUDA platform has been built over nearly two decades into the standard programming environment for AI, so that even a comparable rival chip requires significant re-engineering before anyone can use it. Switching hardware is not simply a matter of buying different hardware.

Beside that ecosystem stands the other durable anchor, which is the hyperscale cloud oligopoly: the small group of companies that operate computing infrastructure at global scale. What sustains their position is capital intensity that few organisations on earth can match. Combined capital spending by these companies is forecast to exceed $600 billion in 2026, most of it directed at AI infrastructure. A sum on that scale is itself the barrier to entry.

So the durable bottleneck, named precisely, is the cloud and the software ecosystem. It is not "compute", and it is not "the chips".

That distinction is not academic. The market for chips is exactly where a defendant will point in order to rebut a claim that computing power is locked up, and any case built on the hardware market will deserve the rebuttal it receives.

Which Law Attaches Where

Once the layers are separated, each one turns out to raise a different legal question.

At the compute layer, the questions concern refusal to deal and the essential facilities doctrine, which together govern when a firm controlling something others genuinely need can be required to supply them. Those arguments work only if they are aimed at the cloud and the ecosystem.

At the model layer, the live issue is the entanglement between AI laboratories and their cloud providers. Several of these arrangements look, on inspection, like mergers that avoided merger review, and their unwinding avoids review a second time on the way out. A later article returns to that problem.

At the agent layer, the territory is defaults, self-preferencing, which means a company favouring its own services over rivals', and control of the gateway through which users reach everything else. This is the search engine and app store playbook, transposed to assistants.

Downstream, ordinary competition law does what it has always done.

Running across all four layers is the fact that redeems the grain of truth in the original sentence. A small number of firms are present at *several layers at once*. So the concern that survives careful analysis is not that one firm controls AI. It is that power held at one layer can be levered into another: control of computing shapes who can build models, model partnerships shape who supplies agents, and control of the gateway shapes which agents reach users.

Why This Matters

The monolith is the source of most bad reasoning in this field, and it misleads in both directions.

Treat the whole stack as locked, and you will urge drastic intervention against layers that competition is handling perfectly well on its own, while pleading a market definition that a competent defendant will dismantle.

Treat the whole stack as contestable, and you will wave away the one or two layers where a durable bottleneck really does sit, conceding a genuine theory of harm before it is even filed.

Taking the stack apart does not dissolve the concentration worry. It states that worry correctly for the first time, and a correctly stated problem is the only kind anyone can litigate.

Food for thought. *For in-house counsel:* at which layer does your company actually sit, and which doctrine, therefore, is the one that reaches you first? *For practitioners:* would your market definition survive the defendant pointing at the opening chip market? *For policymakers:* is your intervention aimed at the layer that is actually locked, or at the one that is loudest?

Next in this series: the engine everyone assumes (the flywheel that supposedly makes AI concentration inevitable), and why it may be a mirage.

Join the conversation. Tell me which layer I have mis-graded, and why: the comments are open, and the best objections will shape later editions.