Bertin Martens

The Incongruent Economics of AI

On Thursday, December 5, Bertin Martens joined Markus' Academy for a conversation on The Incongruent Economics of AI. Bertin Martens is a Senior Fellow at Bruegel.

A few highlights from the discussion.

- **A summary in four bullets**
  - The talk highlighted four key incongruencies in the economics of AI: (1) the tension between the growing fixed costs of AI and the productivity growth it will deliver, (2) new questions in competition policy, (3) the challenge that centralized learning poses for individual data rights, (4) the EU's AI policy dilemmas
  - Even under the most optimistic scenarios investment costs will be a chokepoint for the development of AI. Developers are already working on shifting workloads from upfront investments to operational costs
  - "Co-opetition deals" between startups and big tech foster competition while also reinforcing big tech's market power. Are they necessary to promote AI development?
  - The EU faces several disadvantages in AI, including the absence of major tech companies and high regulatory and electricity costs. Leveraging open-source and "derived" LLMs could help unlock productivity gains

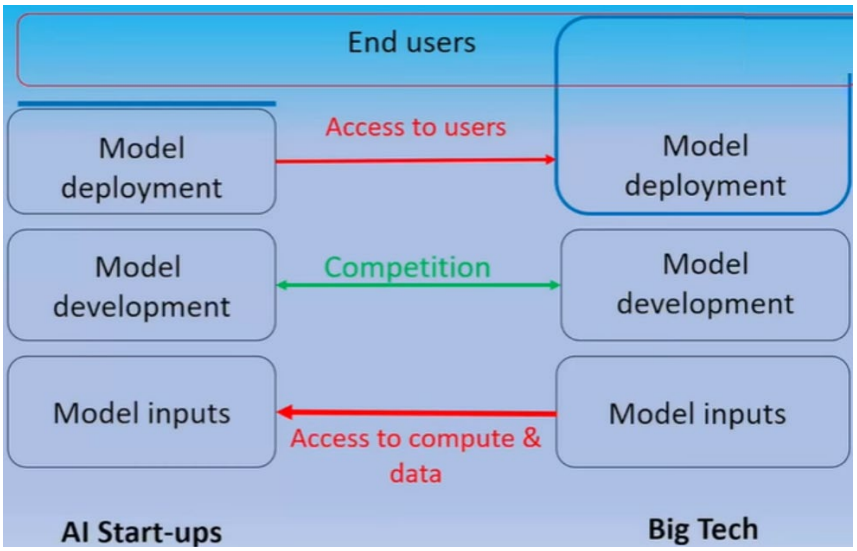
- **[0:00] Markus' introduction**
  - Technological transitions often involve a catch-up effect as economies adjust to a new frontier. Fast-moving innovations can create a "J-curve" effect, where after the arrival of a technology productivity can decline at first. The speed of adoption becomes critical for resilience and to minimize the initial downturn
  - Many open questions remain in the economics of AI. To what extent will AI become systemically interwoven in the economy? Should the creation of future training data be remunerated? Should there be patents on the outputs of LLMs? These may be unnecessary if the sector is already a natural monopoly
  - The EU faces a dilemma as a laggard in AI: strict data protection limits the availability of European training data, potentially preventing EU values and attitudes from being reflected in LLMs. This could lead to a misalignment between the EU's social norms and those embedded in the LLMs it uses

- **[9:52] Exponential cost growth and the productivity growth from AI**
  - AI model training costs are growing at 4-5x per year. From 2016 to 2024 training a model has gone from costing $1,000 to hundreds of millions. These costs are driven by staff (40%), chips (30%), infrastructure (15%), interconnections (10%), and electricity (5%)

    - Computing infrastructure costs reach 10 times the costs of training a model, while chips, with an 8-month depreciation cycle (Cottier et al., 2024), become outdated rapidly
    - In 2024 global investment in AI reached $200bn. Extrapolating current trajectories to the mid-2030s, costs will surpass global GDP, clearly exceeding the financial capacity of big tech
    - Because of AI's scaling laws, data, parameters and computing power are complementary inputs (Kaplan et al., 2020). With the elasticity of substitution between these at around 0.74, there is minimal flexibility for input substitution
    - Improvements in model performance will only lead to revenue growth if this performance translates into productivity growth throughout the economy. Yet the estimates of AI's productivity impact vary widely, from Acemoglu's (2024) 0.7% total productivity gains over the next 10 years to Goldman Sachs' (2023) 1.5% annual productivity growth over the same period
    - In any case the productivity gains will be slow to materialize because of the "J-curve" effect (Brynjolfsson et al., 2021). More immediate productivity gains will come from specialized, smaller, and compressed models (Grootendorst, 2024)
    - A back-of-the-envelope exercise shows that a $1tr AI investment boom is only barely justified if it brings a productivity growth of 1.5% (Martens, 2024)
    - Developers are aware of this investment chokepoint and are shifting workloads to recurrent costs by (1) focusing on improving model performance post-training using proprietary or real-time data, and by (2) enhancing model performance during inference rather than training, thereby improving models' ability to reason, process more context, and accumulate memory
    - It is in this context that real-time and proprietary data (like social media data) becomes crucial and raises competition concerns. Network effects will create monopolistic dynamics similar to the ones we saw with search engines, where the best engines became dominant simply by being the most popular

- **[27:50] Three challenges in competition policy**
    - Firstly, the explosion in fixed costs will force collaboration even among big tech, challenging our traditional notions of regulating monopolies. How can we allow for such collaboration, treating AI as a global natural monopoly, while still preserving competition between AI models?
    - Secondly, how should we regulate monopolistic behavior within layers of the AI stack? There is a near monopoly in chips and an oligopoly in compute infrastructure (Google, Microsoft, Amazon). Training will become less competitive as we hit the data wall and social media data becomes more relevant (this explains why xAI is growing so rapidly)
    - Although no anti-competitive behavior smoking guns have been found yet, there is no easy competition-based solution for these bottlenecks. Even if OpenAI retained all its revenue from ChatGPT and the app store, it would still be insufficient to sustain the company
    - Thirdly, we are seeing collaboration agreements between startups and big tech. Although startups develop and deploy their own models they often share these with big tech because they need their compute, data, and chips

- “Co-opetition deals” (for example the one between Anthropic and Amazon) promote competition in that they allow for startups to enter the market. However they are anti competitive in that they strengthen the market power of big tech



- Are these agreements attempts to circumvent M&A rules, or are they necessary to promote AI development? As Teece (1980, 2020) argues, when markets for complementary inputs fail—especially in complex areas like AI where the costs and benefits from combining inputs are hard to price—vertical integration is preferable as it allows large companies to leverage their economies of scope by reusing inputs
- In this context, is it feasible for the EU to enforce the Digital Markets Act requirement that big tech “gatekeepers” provide open access to their software, such as social media apps, app stores, and browsers? Ensuring seamless integration into the vertical stack without disrupting hardware and software functionality is an important engineering problem

- **[50:35] Growing tension between the private and social value of data**
  - Copyright issues in accessing training data are becoming increasingly important. In the last year the share of copyright owners opting out from their data being used for model training has risen from 25% to 40% (Longpre et al., 2024). In the U.S. there are 2,000 copyright-related cases against AI companies
  - Some AI companies have been making deals with newspapers to obtain their real time data, but for many startups licensing data is unaffordable due to the high transaction costs. These kinds of limitations will reduce model competition and quality, and will increase biases
  - In the EU there is additional regulatory uncertainty on whether it is legitimate to use personal data in model training. Is it compliant with the GDPR to obtain consent only through the general terms of use?
  - Advanced models are being withheld from the EU market due to its data protection regulations, particularly penalizing small language communities

which have insufficient training data from Common Crawl sources and where social media data is necessary for training
    - These issues highlight tensions around the centralization of information, with AI models representing a quantum leap toward further centralization
    - The economies of scale and scope in aggregating and reusing data entail that the social value of data exceeds its private value. As a result, exclusive private rights to data (from privacy and copyright protections) lead to market failures by limiting the social benefits of data sharing
    - The compromise reached with search engines, of allowing only for the extraction of fixed-length strings from articles, is insufficient for AI, which needs access to entire articles for effective training
    - The European Health Data Space demonstrates the value of centralized data, by pooling health data across the EU it is providing clear research benefits

- **[1:01:32] EU AI Policy Options**
    - The EU has several comparative disadvantages in AI. It lacks the computing power infrastructure and a big tech ecosystem, and has high electricity costs
    - It also has regulatory disadvantages from data protection, regulatory uncertainty, and high compliance costs with the AI Act. There may be a reverse “Brussels Effect”, where the EU may have inspired other countries to avoid what it is doing
    - The French Mistral AI is moving to the U.S. due to these factors, along with the desire to tap U.S. capital markets
    - There have been a few policies such as the AI Factories program, EuroHPC to promote supercomputers. The European Commission announced a forthcoming Cloud and AI Regulation to strengthen the position of small providers for them to build large-scale infrastructure
    - However because of these disadvantages the EU cannot afford to invest taxpayer money to reach the AI frontier. Collaboration with US big tech is the only viable option
    - Another option is to thrive below the AI frontier by utilizing open-source models, which require less hardware and computing power for inference, or compressed versions of frontier models that deliver almost the same performance. Smaller firms can obtain large productivity gains from these models
    - Ending the regulatory uncertainty around the AI Act should be a priority, with a focus on clarifying compliance requirements and ensuring a pro-innovation enforcement

- **[1:08:58] Non-economic drivers of AI**
    - The Fear Of Missing Out is driving much of the investment in AI. However, as the dot com bubble showed, even if stock markets pull back the technology will endure, with stabilized growth from the catch up of the J-curve
    - Geopolitical factors, such as China's advancements in AI for both civilian and military applications, are also fueling investment. China's AI models rival those of the U.S. despite lacking access to advanced chips, thanks to their use of vast datasets, including unrestricted social media data

    - Altruistic motivations, such as "AI for good" initiatives, also promote AI development. OpenAI was a key driver of this vision until recently

Timestamps:
[0:00] Markus' introduction
[9:52] Exponential cost growth and the productivity growth from AI
[27:50] Three challenges in competition policy
[50:35] Growing tension between the private and social value of data
[1:01:32] EU AI Policy Options
[1:08:58] Non-economic drivers of AI