Imperial College Business School: ESB Report
Strategy & Digital Perspectives

Trust, Transparency, Incentives, and Adoption Challenges in AI Business Applications

Word Count Excl. Title, Headings, Figures, References, Acknowledgment: 3,292

Imperial College London
Imperial College Business School
MSc Economics and Strategy for Business
16/08/2024

Table of Contents:

Trust, Transparency, Incentives, and Adoption Challenges in AI Business Applications

ACKNOWLEDGEMENT

The following report is informed by recent personal research and consulting experience at Accenture Strategy on the topic of AI applications to M&A processes, facilitated through the ESB Consulting Project. Recency of information sources is prioritised, owing to rapidly evolving AI market developments and lack of sufficient updated academic literature.

ABSTRACT

As AI technologies advance, businesses face rapidly evolving landscapes filled with opportunities and hidden challenges. These innovations are transforming industries, but they also raise important questions about the transparency and interpretability of AI systems. The use of uninterpretable AI methodologies carries risk due to their lack of clarity in reasoning, potentially undermining trust among users and stakeholders. Simpler "white box" models, though less powerful, can offer greater transparency and may be better suited for contexts requiring high accountability. Choosing between open-source and closed-source AI models further complicates decision-making. Open-source models offer flexibility and innovation but require careful handling of data privacy and specialized expertise. Closed-source models, while offering easier integration and vendor support, may lead to high long-term costs and vendor lock-in. This report explores the advancements and trade-offs in decisions regarding AI investment and data strategies, current developments in the AI industry, historical parallels to prior tech industry trends, and analyses recent examples of AI transparency issues and real-world applications for firms through game theoretic and information economics approaches. It emphasizes the strategic importance of fostering peer-to-peer trust and responsible AI use through robust management practices, which are essential for businesses aiming to fully harness AI's productive potential.

INTRODUCTION & AI MARKET RECAP

Recent advancements in artificial intelligence (AI) have significantly advanced the commercial AI landscape. OpenAI's public release of GPT-3.5 in November 2022 marked a turning point in language processing capabilities, followed by GPT-4 in March 2023, which further pushed the boundaries of AI by handling both visual and linguistic tasks, demonstrating the expanding multimodal versatility of AI applications beyond text processing capabilities. Advances in AI-driven image creation were highlighted by the release of DALL-E 2, an improved model from OpenAI generating detailed images from text descriptions, showcasing the potential of AI in creative industries. Tools like Sonix AI indicate a practical audio application of AI in content processing, enabling automatic transcription and summarization of audio files and video

recordings. Features included in the introduction of models like Anthropic's Claude 2 focused on safer interactions in conversational AI, addressing growing concerns over AI ethics and safety. Google DeepMind's Gemini 1.5 exemplified the future of AI as another model capable of handling diverse tasks across visual and linguistic domains, but also raised transparency concerns regarding developer biases in their AI models.

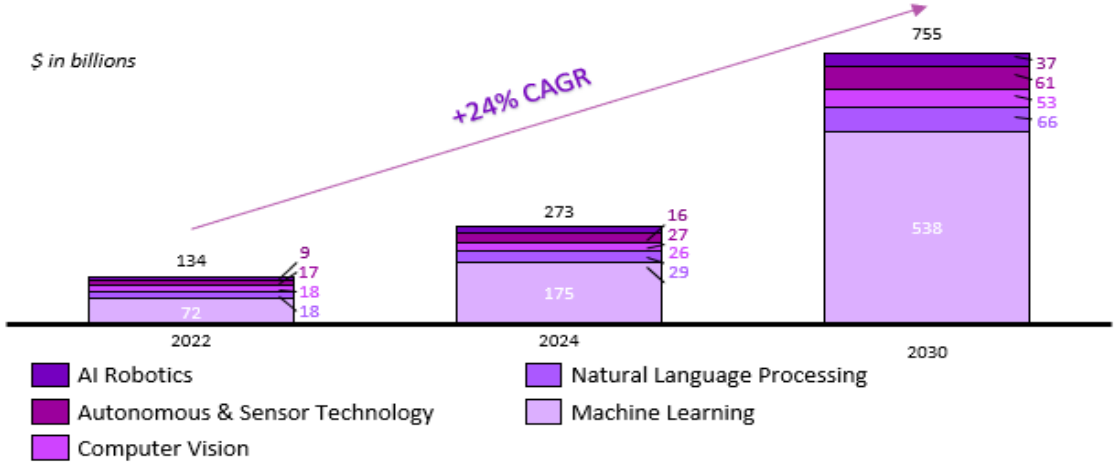

Fig. 1: AI Market Growth Forecast by Application to 2030 (Meridian Capital, 2024)

Looking forward, several trends and future developments are poised to shape the AI industry. Generative AI and smaller language models are expected to continue revolutionizing automated content creation, with expanding applications in sectors like healthcare, legal review and mergers & acquisitions, where consistency as well as speed of execution are paramount (SS&C, 2024). The rise of multimodal AI will enhance output ranges and offer more nuanced insights (IBM, 2024). Open-source AI advancements are anticipated to speed innovation and collaboration, making AI tools more accessible and cost-effective for a broader range of users (ADaSci, 2024). However, the prevalence of unsanctioned AI use within organizations will likely necessitate stronger policies to manage these tools effectively to mitigate consequences from untransparent AI usage (eWeek, 2024).

TRUST & TRANSPARENCY IN AI APPLICATIONS

Trust is emerging as the cornerstone of successful AI adoption. For AI applications to be widely accepted, they must be transparent in their operations. Transparency involves making the decision-making processes of AI systems understandable to stakeholders, administrators, and end users (HBR, 2023). This is crucial for addressing mistrust and scepticism among users. "Black box" models are complex models which are difficult to explain and require great expertise to understand. Methodologies including Neural Networks and Deep Learning offer superior computing performance vs. simpler "white box" models. They excel in highly iterative tasks such as medical diagnostics in radiology, materials composition, pharmaceutical discovery, fraud detection, and security screening, where procedural justice concerns are minimal. On the

contrary, "white box" models are more interpretable, smaller models. Methodologies include Regressions, Decision Trees, and Rules-Based Systems, which are understood by end users and administrators. While these models are less computationally powerful than black-box models, they offer advantages in procedural justice, have lower computational requirements, and offer compellingly similar performance (HBR, 2023). They are proven suitable for complex applications like criminality and loan default prediction, performing within a 1% margin of error of black box models of the same purpose.

Some business processes, such as legal proceedings, consumer banking, and M&A activities requiring a high degree of accountability benefit from the clear trail of reasoning provided by white-box AI models. Comparative accuracy of white-box models indicate a viable path for early adoption of simple AI tools for repetitive tasks in traditional white collar business processes such as document analysis, applying valuation models, and homogenizing data throughput (IMAA, 2023). Pilot teams and management-sanctioned use cases for AI tools can increase trust and uptake by employees. Confirming the expected balance of human decision making and trust in model outputs can create safe boundaries for employees and reduce human error in the application of AI tools. Employee trust in their AI tools underpins successful AI adoption. Trust in black box model conclusions can be developed through trust in the peers and developers responsible for creating models, without need for understanding model machinations (HBS, 2023). Use of simple models entails oversimplification risks, where the clarity of white-box models may incline users to overrule model output with personal judgement, arising from overconfidence in users own perceived understanding of AI insights.

To demonstrate the importance of trust, fashion retailer Tapestry Inc. (owner of luxury brands Coach and Kate Spade) participated in a study on the topic of employee trust in AI tools, where employees responsible for storefront stocking were provided with 2 sets of recommendations by their managers, one from a white-box model and one from a black-box model. Comparisons to the optimum determined allocations were followed for both groups. The results from the study highlighted counter-intuitive findings: the white box allocators, who perceived they understood their recommendation algorithm better, were more likely to overrule their recommendation, resulting in suboptimal allocations, whereas the black box allocators who did not understand their algorithms still followed the recommendation, performing 26% better in the study. This outcome arose from "social-proofing" the algorithm, where the black-box allocators had placed their trust in their immediate peers who had helped to test and develop the model, and using this peer-to-peer trust as a foundation had followed through on recommendations which they did not understand, subsequently achieving more optimum results (DeStefano et al., 2022).

OPEN VS CLOSED SOURCE AI APPROACHES

For many firms, approaches to inputted data will determine the success of their AI insights. Different approaches to data collection offer unique advantages across sources, spanning freely available public data, to costly private data products, to arduous primary collection. Different model approaches offer unique advantages and trade-offs, with ease of integration and expertise required for implementation varying widely across model approaches. Choosing an AI solution that fits to the tailored needs of specific industries and business processes is highly subjective and requires careful consideration. To proceed with this topic, a discussion of the characteristics of open and closed source models and data inputs is warranted.

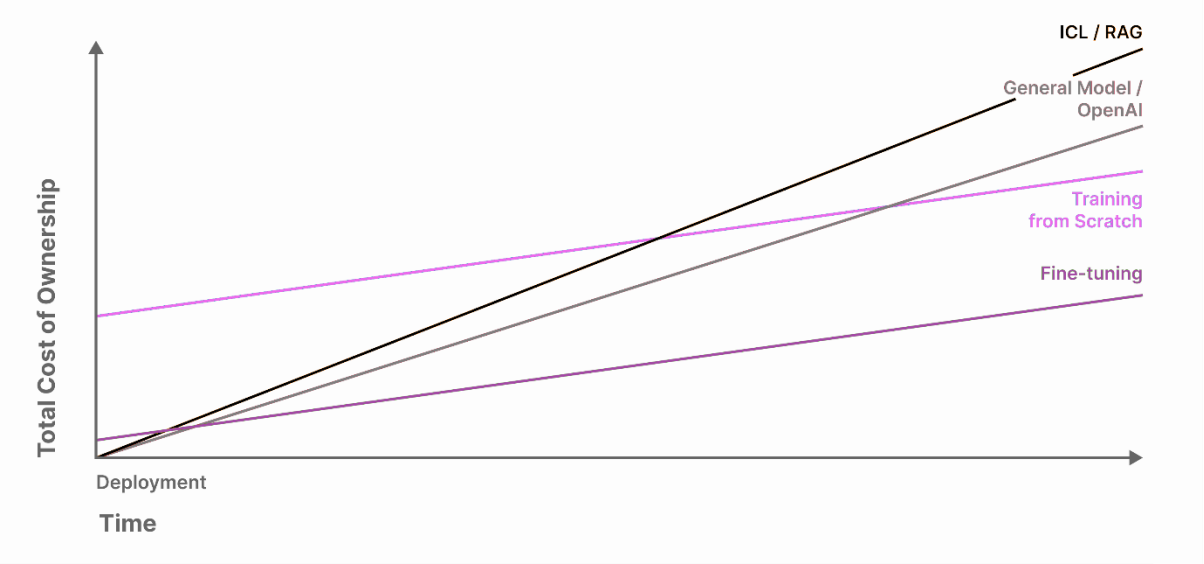

Fig. 2: Deployment Costs of Different Model Approaches over Time (Predibase, 2023)

[ICL = In-Context Learning, RAG = Retrieval-Augmented Generation]

[Both are methodologies for training models on very little data, often for zero-shot never-before-seen tasks. Used in situations where little to no training data exists.]

Open-source data offers significant advantages for simple AI applications. Publicly accessible datasets are often extensive enough to train most NLP programs to analyze text effectively. The community-driven nature of open-source datasets fosters continuous innovation, peer review, and transparency. Open-source data encompasses a broader domain than proprietary data, allowing for more diverse training (ADaSci, 2024). However, the generality of publicly available data may not align with specific training needs, necessitating additional efforts to tailor data for niche use cases. Varying formats and quality of public data require more rigorous cleaning and standardization before it can be used effectively in model training or fine-tuning. There are also concerns related to regulatory compliance, particularly regarding the licensing, processing, and storage of public user data for commercial use (Attri, 2024).

When considering open-source models, the benefits are notable. These models are typically free to use, though some require case-by-case commercial licensing, and offer a cost-effective solution at scale (see fig. 2). Open-source models can be fine-tuned to meet specific requirements using confidential data, which enhances their adaptability in niche use cases. The active open-source community supports these models, leading to rapid innovation, increased transparency, and many diverse applications (MIT Technology Review, 2024). On the downside, the complexity involved in fine-tuning open-source models can present challenges for unexperienced staff, requiring additional technical training or the hiring of AI specialists. Data privacy concerns are acute with open-source models as the responsibility for cybersecurity and regulatory compliance shifts onto the administrators, as opposed to relying on vendor-provided services (CIC, 2023). This increased responsibility extends to both the use of the model and the management of the input data, making it crucial for organizations to weigh these factors carefully when considering open-source AI solutions.

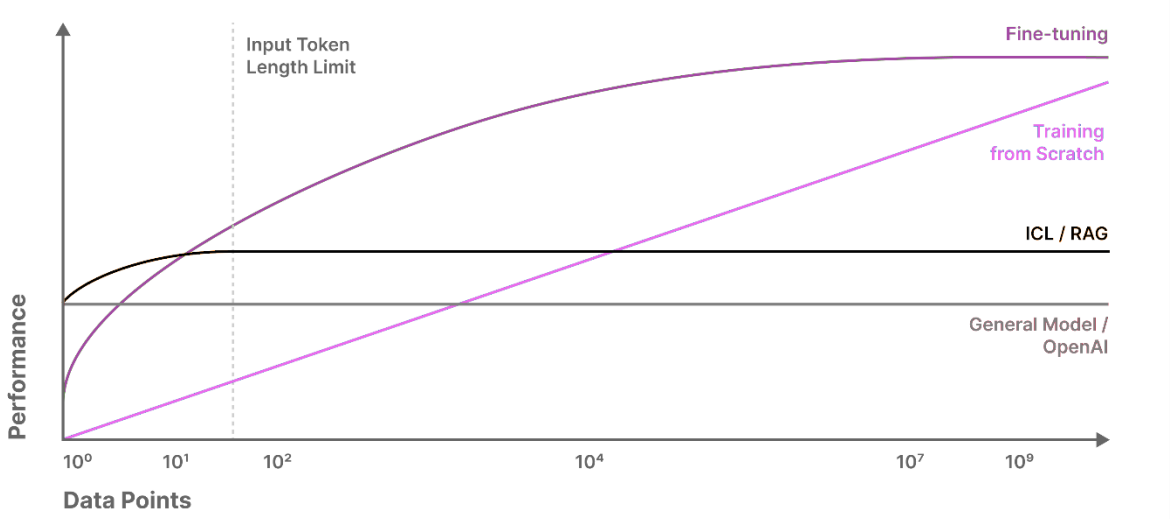

Fig. 3: Performance of Different Model Approaches vs Input Data (Predibase, 2023)

In parallel, closed-source data can provide exclusive insights. The use of private data can grant organizations a competitive edge through first-mover advantages in utilizing data not available in the public domain (HBR, 2013). Closed-source data can be instrumental in fine-tuning open-source models for unique use cases (see fig. 3), enabling a more customized and cost-effective application of AI technologies. However, there are disadvantages to closed-source data. Costs associated with purchasing private data or collecting primary data can be significant, which may be a barrier for some organizations. Privacy and transparency concerns are more pronounced with proprietary data, particularly regarding confidentiality and ethical use. The potential for bias within small datasets is a substantial challenge, often necessitating de-biasing, sandboxing, and isolation practices to ensure the integrity of resulting insights.

When it comes to closed-source models, the benefits are compelling. These models often integrate seamlessly with pre-existing ecosystems, thanks to vendor-provided support and simple API interfaces (TechTarget, 2023). The dedicated support offered by vendors ensures that models are tuned to specific requirements, with the vendor also taking responsibility for security, performance, and optimization updates. Additionally, vendors such as Amazon Web Services often provide the necessary computing assets or cloud-based solutions, reducing the burden on the organization's infrastructure. However, the use of closed-source models is not without its drawbacks. The costs associated with scaling these models can become prohibitive, particularly if an organization becomes reliant on a single vendor, leading to vendor lock-in. Furthermore, the use of shared vendor infrastructure can result in rate limits, quotas, and potential disruptions during periods of high traffic (Attri, 2024). Privacy concerns arise as vendors may use private data to train their own models, necessitating strict privacy agreements and confidentiality due diligence.

CURRENT & HISTORICAL AI MARKET DEVELOPMENTS

Companies like OpenAI, Microsoft, Anthropic, and Meta are grappling with the evolving demands of innovation, competition, and financial viability. Reports suggest that OpenAI could be on the brink of bankruptcy within the next 12 months, with projected losses amounting to up to $5 billion (ITPro, 2024). These financial difficulties reflect the high operational costs of developing and maintaining frontier AI models like GPT-4, coupled with increasing competition. The financial strain on OpenAI underscores the challenges of monetizing closed source AI technologies at scale and balancing heavy capital expenditure. Despite pioneering advancements, OpenAI's business model, which relies heavily on cloud services, user subscriptions, and API usage, has yet to demonstrate long-term profitability. This uncertainty raises questions about the sustainability of commercial AI development,

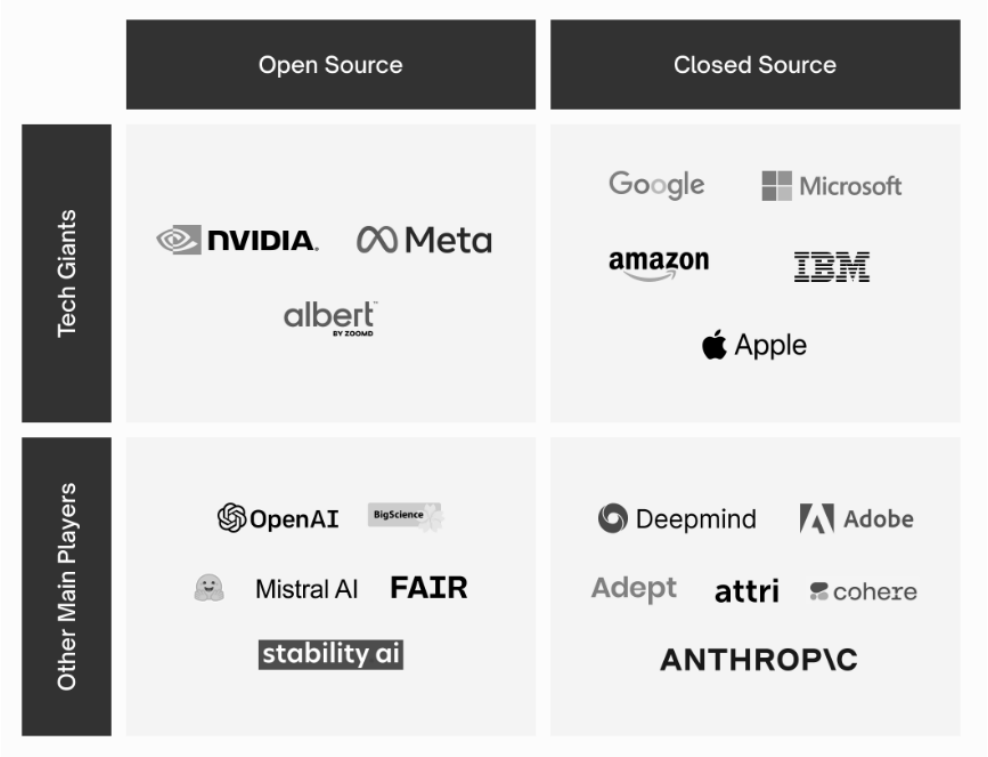

Fig 4: Dominant AI Market Players by Size and Access Approach (Attri, 2024)

To understand the current closed-source vs. open-source debate in AI, it's informative to look back at the Windows vs. Linux "server wars" of the late 20th and early 21st centuries. During this period, Microsoft dominated the enterprise software market with its proprietary Windows operating system and suites of office software, while Linux emerged as a powerful open-source alternative. The conflict between these two models highlighted fundamental differences in business strategy: Microsoft's focus on proprietary software and market control versus Linux's emphasis on community-driven development and open access principles. In the end, both approaches found their place in the industry, with Linux's low costs and community support key to its continued success (Volico, 2024). While Microsoft maintained its dominance in the consumer and enterprise markets, Linux became the backbone of most server environments and the foundation of numerous computing technological innovations, including the cloud

computing infrastructure that powers much of today's internet, eventually leading to Microsoft conceding open access to some 60,000 Linux-related patents (Wired, 2018).

Today, a similar debate is unfolding in the AI domain. OpenAI, backed by Microsoft, represents the closed-source approach. Proprietary models allow for controlled development and a more captive user base, but raises concerns about the concentration of power, shared vendor infrastructure, and accessibility of AI technology. Meta is embracing a more open approach to AI, making significant strides in its open-source AI models, namely the LLaMA series. Meta's strategy is to "corner users" by democratizing AI tools, integrating them into existing Meta products, and allowing a broader range of developers to build upon Meta's work while firmly asserting Meta as the main provider of consumer AI interfaces (Business Insider, 2024). This approach mirrors the Linux philosophy and could lead to widespread adoption, particularly in areas where proprietary models might become too expensive or restrictive (L'Atelier BNP Paribas, 2023). The ongoing debate between closed-source and open-source AI is not just about technology—it's about control, accessibility, trust, and the future of innovation. Just as the Windows vs. Linux server wars eventually led to a hybrid landscape where both proprietary and open-source software coexist for differing purposes, the AI industry might be heading towards a similar equilibrium. OpenAI's financial challenges highlight the risks of a closed-source business model that relies heavily on monetization through restrictive access. In contrast, Meta's open-source strategy could foster a larger and more inclusive AI ecosystem in the long-term, but faces the challenge of monetizing open-source technology without the direct control that comes with proprietary systems.

ECONOMIC APPROACHES TO UNDERSTANDING AI MARKET DEVELOPMENTS

For the firms building the AI models, substantial infrastructure expenditure is required to train and scale AI models. This has led to a dynamic resembling a game of escalation in AI capital expenditure, in which seemingly irrational outcomes of overbuilding begin to make more sense once all incentives are considered.

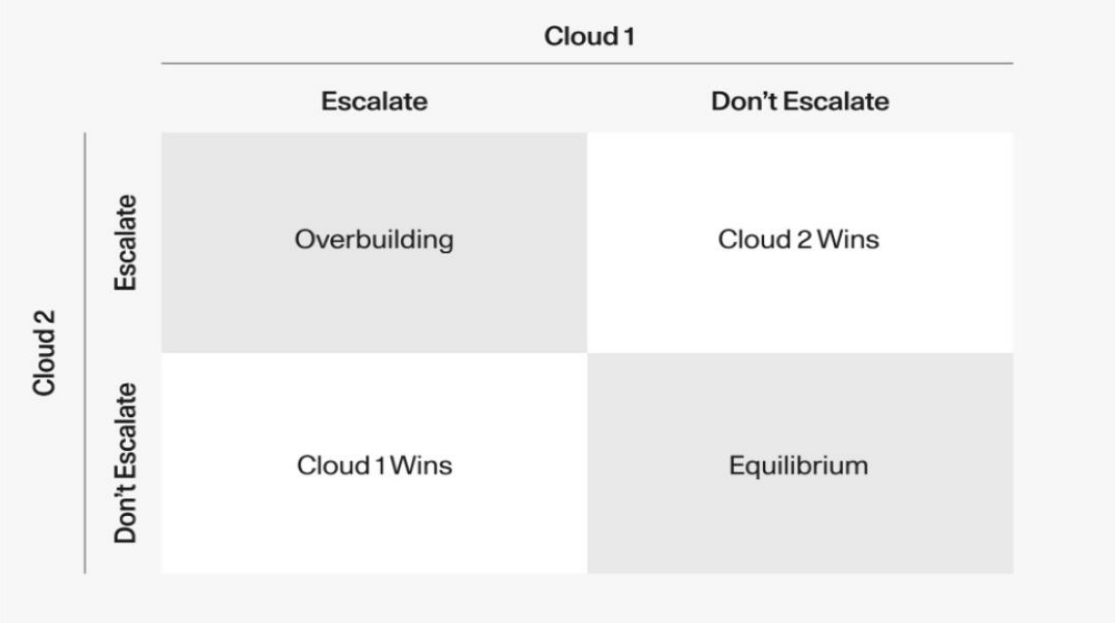

Fig. 5: AI CapEx Escalation Game (Sequoia Capital, 2024)

At first glance, heavy initial AI capital expenditure commitment seems counter-intuitive: GPU architectures and model cluster sizes are evolving at such rapid pace that any present investment choices may become quickly outdated. In a market best described as an oligopoly, scarcity of semiconductors and ballooning frontier compute requirements have created incentives which push industry titans to buy up any available GPUs and vacant data centre real estate for fear of ceding already-scarce resources to competitors, with the added externality of reinforcing supply constraints and erecting extremely costly barriers to entry for smaller firms, essentially restricting participation in the AI infrastructure market to only the richest of players (Sequoia Capital, 2024). In this regard, overbuilding infrastructure early as a defensive tactic against established players and AI startups is a rational strategy for the incumbent AI titans, and will likely continue. This may not be bad for the AI market in the long run, as this fierce competition is well-poised to bring down compute and API usage costs in the AI industry over time.

As machine learning capabilities become more proliferated, they begin to challenge traditional economic relationships of information. A notable example is that of principal-agent problems; is an AI algorithm able to be considered an agent? If so, can some form of "contract" be imposed to remedy the moral hazard arising from delegation of responsibility to a model? In the event AI causes negative outcomes, which party will bear the responsibility of failure? Such problems can be classified as "second principal-agent problems", with the agent being represented by a "superintelligent system", as opposed to "first principal-agent problems" which occur between human parties (Borstrom, 2014). In contexts like automated trading of financial instruments, increasing deployment of machine learning algorithms can create situations where it does not matter to the human principal if the agent model is verifiably following their established contract or not, so long as it is generating profit. As greater responsibility is delegated to machine learning models, rules may no longer be set by humans but instead developed internally by the AI algorithm for itself, further reducing transparency and increasing moral hazard. Where competitors are using the same AI strategies, there is an associated risk of exacerbating negative market fluctuations, or "flash crashes", owing to "resonance" between overly similar trading methodologies and source data (Borch, 2022). Predicting this kind of phenomena becomes more difficult as opacity in trading reasoning increases. The fundamental problem is that we cannot fully interpret deep learning and neural network reasoning, which makes designing effective contracts to specify desired behaviour from AI models exceedingly difficult. This represents a primary challenge to the adoption of AI by businesses, as the inability to explain model outputs may result in material consequence.

Opacity in machine learning model decision making has already resulted in legal and financial repercussions. In 2019, Apple and Goldman Sachs issued a credit card, the Apple Card, and soon found that many women's cards were given up to 20 times lower credit limits than their husbands, despite having better jobs and credit scores. The credit limit allocation was decided by a "black-box machine learning algorithm" and was not interpretable to Apple or Goldman Sachs employees. One of the women impacted was Janet Hill, the wife of Apple co-founder Steve Wozniak, casting a streak of irony through the situation. This gendered variability in credit allocation prompted an investigation by the New York Department of Financial Service into gender discrimination, intentional or otherwise, leading to legal proceedings under the Equal Credit Opportunity Act (AI Incident Database, 2019). This incident highlighted the risk of being unable to explain decisions to regulators when following conclusions from a model which humans cannot easily rationalize, which can alienate customers and result in legal consequence. It is important to note how the accused firms in question were still held responsible for the

unintended consequences of their model, despite no intentional discrimination on the part of Apple or Goldman Sachs. This exemplifies a cautionary tale for companies who rush to adopt AI in business processes without strategic assessment of potential risks owing to information asymmetry between customers, regulators, and administrators of AI models without a plan for mitigating unintended AI outcomes. While incentives may exist regarding speed-to-market, the substantial consequences of AI misapplication should give pause to firms and further emphasize the necessity for stringent risk assessment of AI in specific use cases.

CONCLUSION & RECOMMENDATION

Rapid advancements in AI technologies present both exciting opportunities and significant challenges that require careful consideration by businesses. While "black box" models may offer nominally superior performance, their lack of interpretability can lead to mistrust, legal challenges, and reputational risks. Organizations must weigh the benefits of these advanced models against the potential risks, particularly in contexts where accountability and procedural justice are important. By defaulting to interpretable AI models where feasible, businesses can mitigate risks associated with opaque decision-making and build stronger trust among stakeholders, including customers, regulators, and employees. The debate between open-source and closed-source AI models underscores the importance of strategic planning in AI adoption. Open-source models, while fostering innovation and cost-efficiency, require robust data privacy measures and specialized expertise. In contrast, closed-source models offer ease of integration and vendor support but may result in long-term dependencies and higher costs.

Building trust within organizations emerges as paramount. The success of AI adoption hinges on the trust employees place in their AI tools and the peers who develop and manage these systems. Pilot teams and clear guidelines from management level can further cement trust in AI tools. Ultimately, the key to successful AI integration lies in balancing innovation with responsibility. By adopting interpretable models where possible, fostering peer-to-peer trust from development to execution, and implementing strong management guidelines, businesses can leverage AI to its fullest potential while minimizing risks. This approach will ensure that AI technologies are not only powerful tools for growth, but also trusted partners in the journey toward sustainable and ethical business practices.

References:

ADaSci (2024). *Proprietary VS Open-Source AI Models in Generative AI.* Available at: https://adasci.org/proprietary-vs-open-source-ai-models-in-generative-ai/ (Accessed: 12th August 2024)

AI Incident Database (2019). *Incident 92: Apple Card's Credit Assessment Algorithm Allegedly Discriminated against Women*. Available at: https://incidentdatabase.ai/cite/92/ (Accessed: 15th August 2024)

Attri (2024). *Choosing Your Path in Generative AI: Open-Source or Proprietary?* Available at: https://attri.ai/blog/choosing-your-path-in-generative-ai-open-source-or-proprietary (Accessed: 12th August 2024)

Borch, C. (2022). Machine learning, knowledge risk, and principal-agent problems in automated trading. *Technology in Science [68]*

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. *Oxford University Press.*

Business Insider (2024). *Mark Zuckerberg has brilliantly cornered us.* Available at: https://www.businessinsider.com/mark-zuckerberg-meta-wins-ai-race-by-cornering-users-2024-8 (Accessed: 14th August 2024)

CIC (2024). *5 investors on the pros and cons of open source AI business models.* Available at: https://www.cic.vc/5-investors-on-the-pros-and-cons-of-open-source-ai-business-models/ (Accessed: 15th August 2024)

DeStefano, T., Kellogg, K., Menietti, M. and Vendraminelli, L. (2022). Why Providing Humans with Interpretable Algorithms May, Counterintuitively, Lead to Lower Decision-making Performance. *SSRN Electronic Journal*.

eWeek (2024). *10 Most Impactful AI Trends in 2024.* Available at: https://www.eweek.com/artificial-intelligence/ai-trends/ (Accessed: 13th August 2024)

Harvard Business Review (2023). *AI Can Be Both Accurate and Transparent.* Available at: https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent (Accessed: 14th August 2024)

Harvard Business Review (2013). *Invest in Proprietary Data for Competitive Advantage.* Available at: https://hbr.org/2013/03/invest-in-proprietary-data-for (Accessed: 12th August 2024)

Harvard Business School Working Knowledge (2023). *What Makes Employees Trust (vs. Second-Guess) AI?* Available at: https://hbswk.hbs.edu/item/what-makes-employees-trust-vs-second-guess-ai

IBM (2024). *Data Suggests Growth in Enterprise Adoption of AI is Due to Widespread Deployment by Early Adopters, But Barriers Keep 40% in the Exploration and Experimentation Phases.* Available at: https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters (Accessed: 10th August 2024)

IMAA (2023*). Transforming the M&A Process: The Current and Future Role of Artificial Intelligence.* Available at: https://imaa-institute.org/blog/the-future-role-of-artificial-intelligence-in-mergers-and-acquisitions/ (Accessed: 14th August 2024)

ITPro (2024). *OpenAI could go bankrupt in 12 months if it doesn't raise some serious cash – but is the Microsoft-backed AI giant too big to fail?* Available at: https://www.itpro.com/technology/artificial-intelligence/openai-could-go-bankrupt-in-12-months-if-it-doesnt-raise-some-serious-cash-but-is-the-microsoft-backed-ai-giant-too-big-to-fail (Accessed: 15th August 2024)

L'Atelier BNP Paribas (2023). *Everything you need to evaluate open-source (vs. closed-source) LLMs.* Available at: https://atelier.net/insights/evaluating-open-source-large-language-models (Accessed: 15th August 2024)

Meridian Capital (2024). *Artificial Intelligence M&A Trends: Spring 2024.* Available at: https://meridianib.com/ma-trends/artificial-intelligence-ma-trends-spring-2024/ (Accessed: 13th August 2024)

MIT Technology review (2024). *The tech industry can't agree on what open-source AI means. That's a problem.* Available at: https://www.technologyreview.com/2024/03/25/1090111/tech-industry-open-source-ai-definition-problem/ (Accessed: 15th August 2024)

Predibase (2023). *The Future of AI is Specialized.* Available at: https://predibase.com/blog/the-future-of-ai-is-specialized (Accessed: 14th August 2024)

Seqouia Capital (2024). *The Game Theory of AI CapEx.* Available at: https://www.sequoiacap.com/article/ai-optimism-vs-ai-arms-race/ (Accessed: 14th August 2024)

SS&C (2024). *2024 SS&C Intralinks Dealmakers Sentiment Report.* Available at: https://www.intralinks.com/resources/publications/2024-ssc-intralinks-dealmakers-sentiment-report (Accessed: 13th August 2024)

TechTarget (2024). *Attributes of open vs. closed AI explained.* Available at: https://www.techtarget.com/searchenterpriseai/feature/Attributes-of-open-vs-closed-AI-explained (Accessed: 14th August 2024)

Volico (2024). *Linux or Windows Servers? What's the Difference and Which One's Better?* Available at: https://www.volico.com/linux-or-windows-servers-whats-the-difference-and-which-ones-better/ (Accessed: 14th August 2024)

Wired (2018). *Microsoft Calls a Truce in the Linux Patent Wars.* Available at: https://www.wired.com/story/microsoft-calls-truce-in-linux-patent-wars/ (Accessed: 14th August 2024)