UCLouvain

UNIVERSITÉ DE NAMUR

Economics School of Louvain - ESL

Economics School of Namur - ESN

From access to parity — Can the DMA secure effective interoperability for on-device AI agents?

Theory and evidence from the mobile AI stack

Author : Antonin Fonck

Thesis Director : Johannes Johnen

Thesis Reader : Alexandre de Streel

Academic Year 2025-2026

Master in Economics – 120 credits – Focus : ECON2 MS/G

Université catholique de Louvain - Economics School of Louvain - Place Montesquieu 3, 1348 Louvain-La-Neuve
Université de Namur – Economics School of Namur – Rempart de la Vierge 8, 5000 Namur

Acknowledgements

I would like to thank my thesis supervisor, Professor Johannes Johnen, for his guidance throughout this work. His comments shaped the economic model at its core, from the choice of instrument to the structure of the game, helping improve the argument at every stage.

I am grateful to Professor Alexandre de Streel for agreeing to read this thesis, and for a body of work on the Digital Markets Act on which the legal analysis draws throughout.

My thanks also go to the professors and staff of the Economics School of Louvain and the Economics School of Namur, whose teaching over these two years provided the tools this thesis puts to use.

I thank my brother for reading this thesis and for his comments on it, and my family and friends for their support throughout its writing.

Generative AI tools were used in the preparation of this thesis as a support for language editing, for the formatting of LaTeX code, and for structuring the argument. The scientific work and all substantive claims are my own.

Abstract

The Digital Markets Act (DMA) was conceived before the rise of large language models (LLMS). It is therefore worth asking whether its obligations reach the new competitive bottleneck that AI agents have created on the device itself. Taking the mobile market as its focal case, this thesis asks whether Article 6(7) DMA on interoperability, with Article 6(5) DMA on self-preferencing, effectively reaches a gatekeeper's privileged access to the on-device AI accelerator (the neural processing unit) for its own agent, or whether the obligation's access-not-parity scope, its integrity exception, and its enforcement mode leave a hardware-layer foreclosure gap. It combines a doctrinal reading of Article 6(7) and its enforcement, a reduced-form model of foreclosure by access-quality degradation, and a case study of the Commission's Google proceeding (Case DMA.100220) and of Apple's June 2026 decision to withhold its rebuilt Siri from iPhones and iPads in the European Union.

Keywords: Digital Markets Act; interoperability; Article 6(7); on-device AI; neural processing unit; self-preferencing; foreclosure; AI agents; Apple; Google.

Contents

1 Introduction

Artificial intelligence is at an inflection point. In the space of a few years the dominant mode of consumer AI has shifted from systems that generate text on request to agents that plan, remember, and act (booking, purchasing, and operating other applications on a user's behalf), and the frontier is still moving quickly, with competing laboratories releasing successive and markedly more capable model generations at a pace the law was not designed to track. With devices now able to run AI locally, the integration of agentic AI is accelerating, and an on-device agent's reach is expanding with it: to a user's biometric authentication, to other applications, to on-screen context, and to personal information held on the device. What agentic AI can ultimately do for a user is less and less limited by the raw quality of the underlying model. The binding constraint is increasingly shifting towards the access it is granted: to the user's data, to other applications, and to the hardware of the device. This access is becoming central to how agents will interact with mobile devices, and with agents increasingly mediating the user's interaction with the market (choosing which app to open, which service to call, which product to surface, which transaction to complete) whoever controls the agent shapes how hundreds of millions of users discover and reach the businesses on the other side of the device. The agent is becoming a principal interface to the digital economy, and control of that interface is control of a market. As capability migrates toward access, control of access becomes a decisive competitive asset.

On the agentic AI segment, that shift changes who holds power, and it does not always favour the firm with the best model. The firms that control the most valuable access points are not necessarily the firms at the AI frontier. A gatekeeper in that position faces an obvious temptation: rather than compete on the quality of its agent it can compete on its control of the device, granting its own agent privileged access to the hardware while restricting what a rival's agent may do with the same hardware.

This is what makes the concern acute. The mobile platform layer is effectively a duopoly: as of 2026, Android and iOS together account for almost the whole market, holding roughly 70% and 30% of mobile operating systems worldwide, respectively (StatCounter, 2026). A single gatekeeper could therefore come to foreclose the on-device agent market on its own OS and grant its own agent privileged access to the slot reserved for agentic AI while degrading what a rival agent may do. With a market this concentrated, a firm restricting access to its on-device AI, granting its own agent favourable treatment and in turn using that agent to tilt the downstream markets affects millions of consumers.

The European Union has an instrument that appears to anticipate exactly this. The Digital Markets Act obliges a designated gatekeeper, under Article 6(7), to grant business users and providers of services effective interoperability with, and access for the purposes of interoperability to, the same hardware and software features that are available to the gatekeeper's own services, where those features are accessed or controlled via the operating system (Regula-

tion (EU) 2022/1925). Because the on-device accelerator (the neural processing unit (NPU) on which an agent's inference runs) is reachable only through interfaces the operating system controls, it falls within the provision's scope (Regulation (EU) 2022/1925). The harder question lies one step beyond the text: whether an obligation that secures access secures it on quality-equivalent terms in practice.

The question is no longer hypothetical. In April 2026 the European Commission adopted preliminary findings against Alphabet in the first Article 6(7) proceeding to reach on-device AI (Case DMA.100220[1]), and in June 2026 Apple announced that its rebuilt agentic assistant, Siri AI, would not ship to iPhone and iPad users in the European Union, attributing the decision to the DMA.

The thesis proceeds by three complementary methods. First, a doctrinal reading of Article 6(7), alongside its preliminary findings in Case DMA.100220. Second, a reduced-form economic model: a three-stage monopolistic device-seller game, solved by backward induction, with a quality-of-access choice variable. Third, a case study of the Google proceeding and of Apple's withholding of Siri AI, drawing on the Commission's and Apple's public statements.

The literature chapter reviews the relevant work on foreclosure, self-preferencing, interoperability, and the emerging economics of AI agents. A technical chapter fixes the vocabulary the argument needs: inference and training, agentic and on-device AI, the assistant slot, and the chip-level resources (NPU, random access memory (RAM) and residency, background execution, central processing unit (CPU) and graphics processing unit (GPU)) over which access is contested; it closes on the access-parity distinction the rest of the thesis turns on. The baseline model formalises the incentive to degrade a rival's access once outright denial is unavailable, and identifies the conditions under which the gatekeeper settles at the regulatory floor. The final substantive chapter reads Article 6(7) against its enforcement, applying the framework first to Case DMA.100220 and then to Apple, where the same logic is extended to the decision to withhold. A discussion and a conclusion draw the implications together.

The analysis yields three main results. First, reaching the hardware is not securing parity on it: the remedy leaves the gatekeeper's preferential access in place by default, removable only by an affirmative user action, and rests on a non-discrimination standard whose violation is hardly observable by outside parties. Second, the gatekeeper's choice of access quality has no interior optimum: it sits either at full parity or exactly at the regulatory floor, so the legal definition of effective interoperability becomes the market outcome whenever per-user monetisation exceeds a threshold that is itself increasing in the floor. Raising the standard is therefore self-reinforcing

[1]The analysis of Case DMA.100220 in this thesis is stated as of its research cut-off. It rests on the Commission's preliminary findings and proposed measures of 27 April 2026, which were the operative documents when the research was carried out. On 16 July 2026 the Commission adopted its final decision specifying the measures. The measures themselves have been published, but the non-confidential version of the decision, which sets out the Commission's reasoning rather than only the measures it imposes, had not been released when this research closed. Several points changed between the two texts, most notably the user-toggle provision examined in Section 5.2.1, which does not appear in the final measures. Section 6 sets out some of those changes. Statements in the present tense about the content of the remedy should accordingly be read as describing the April draft.

rather than merely incremental. The same analysis produces withdrawal and withdrawal is informative, since a gatekeeper rationally withholds only if it expects to retain users while offering them nothing, so that Apple's withdrawal is simultaneously a cost of the DMA and evidence for its premise. Third, Articles 6(7) and 6(5) are complements rather than substitutes: enforcing the downstream prohibition lowers the monetisation that drives upstream degradation, and enforcing the upstream duty shrinks the captive base that downstream steering exploits.

2 Related literature

2.1 Foreclosure: from denial to degradation

The relevant literature for this work sits at the intersection of two bodies of work: the economics of vertical foreclosure, self-preferencing, and interoperability under the Digital Markets Act (DMA) and an emerging strand on how Artificial Intelligence (AI) reshapes platform competition. Read as a journey up a river, this review starts at the mouth, where the user meets the market, and works its way back to the source, where the hardware is allocated.

Figure 1: Structure of the literature review, read from the agent's output back to the chip it runs on. Each stretch supplies one element of the argument that follows: what holding a user is worth, at the mouth; how a gatekeeper comes to hold the assistant slot, in the middle reaches; and the terms on which a rival reaches the accelerator, at the source. The foreclosure mechanism set out in Section 2.1 runs the whole length.

Before setting off, the current itself: what makes a vertically integrated firm want to degrade a rival's access, at any point along the way. A vertically integrated gatekeeper can entrench its position by controlling a bottleneck and degrading rivals' access to it, (Salop & Scheffman,

1983). Where the tradition asked whether a dominant firm would price below cost to drive out rivals, Salop and Scheffman showed that a dominant firm facing a competitive fringe can do better by raising the fringe's costs: if rivals' marginal or average costs rise, their output contracts and the market price rises, and as long as the dominant firm's own costs rise by less than it gains, the strategy is profitable. Exclusion, on this account, is a matter of degree rather than a binary refusal to deal, and it can be achieved through any lever that handicaps a rival. The "cost" it has in mind is a monetary input cost rather than a degradation of service quality; but its central intuition transfers directly to our framework, and it frames the entire argument that follows.

Two platform-era models, rebut the recurring claim that an integrated firm has no rational reason to foreclose. Padilla, Perkins, and Piccolo (2022) construct a two-period model of a gatekeeper that both sells a device and supplies a complementary service. Through their two-period model the authors are able to demonstrate that where a gatekeeper device seller facing potentially saturated demand has both the incentive and the possibility to foreclose access to rival firms, there is a time inconsistency problem. In the first period the firm would like to promise open access to rival firms on their devices in order to capture the maximum amount of consumers upfront. But once sales slow down on the device market and the firm has a captive installed base, it prefers to exploit that base. Foreclosure becomes more attractive precisely as device-demand growth slows.

Johnen and Shekhar (2026) sharpen the diagnosis by detailing the kind of network effect at stake. The foreclosure incentive is not just about device-demand growth but about the type of network effect, which can be product-specific or ecosystem-specific. A product-specific network effect will only benefit the product, without producing any spillovers to other services, meaning that a rival's growth directly undermines the platform's own product value. Whereas with an ecosystem-specific network effect, the rival adds value to the whole ecosystem, and hence indirectly to the platform's own product, these effects dampen the foreclosure incentives. It formalises Motta's (2023) argument that self-preferencing is a species of vertical foreclosure. Surveying high-profile abuse investigations he reframes the signature digital-platform practices, self-preferencing and the denial or degradation of interoperability, as forms of foreclosure in vertically related or complementary markets. Hagiu, Teh, and Wright (2022) supply the welfare counterweight. They model a dual-mode platform that both hosts third-party sellers and sells its own products, and conclude that an outright ban on the dual mode can lower consumer surplus and total welfare even when the platform would otherwise imitate rivals and self-preference, because its own product also adds competition and variety.

Rey and Tirole (2007), in what remains the canonical synthesis of the foreclosure literature, locate the incentive in a problem of commitment: having sold access to one downstream firm, the owner of a bottleneck input cannot resist selling to another, and the anticipation of that opportunism dissipates the rent it is trying to protect. Denial of access, exclusive dealing and vertical integration are then ways of committing not to deal, and thereby of recovering a profit

that would otherwise be lost to competition among the downstream firms.

Katz and Shapiro (1985), treating the economics of networks, considered compatibility as a strategic choice rather than a technical given: a firm decides how far its product will interoperate with others, weighing the loss of differentiation against the gain in network value. Economides (1996) draws the sharper implication for a firm that owns a platform: it may rationally invite entry into complementary segments rather than foreclose it, because complementarity and network externalities mean that a richer ecosystem of complements raises the value of its own product, part of which it can capture. A device manufacturer stands in exactly this position with respect to the AI agents that run on its hardware. Every user of a rival AI agent is a user it does not monetise, but a better rival AI agent also makes the device itself worth more, the two incentives coexist.

Two features of the foreclosure tradition differ with the setting studied here. Its object is typically access itself, since the bottleneck owner denies an essential input and the question is when denial is rational, whereas Article 6(7) DMA removes denial from the gatekeeper's choice set and leaves only the question of how well a rival is served. What remains once denial is unavailable is degradation, and the literature that anticipates it was developed for regulated network industries, where refusal was already unlawful and quality was the residual margin. Economides (1998) shows that a vertically integrated input monopolist has an incentive to discriminate against downstream rivals in non-price terms, degrading the quality of the input it is obliged to supply rather than its price. It is the quality-side counterpart of the raising-rivals' costs logic set out earlier (Salop "&" Scheffman, 1983). That literature, however, was built for settings regulated on price and interconnection, and it has no counterpart to a legal floor on the quality of access, which is what Article 6(7) introduces when it requires interoperability to be "effective".

Fletcher et al. (2024) provide the strongest justification for that legislative choice of the Digital Markets Act (Regulation (EU) 2022/1925) to be codified as ex-ante obligations: in digital markets entry is often blockaded rather than merely deterred, because concentration results from forces such as scale and network effects rather than only from strategic conduct. This paper reads that economics sits at the heart of the DMA's *objectives and design*, but the DMA deliberately prioritises clarity, speed, administrability and enforceability therefore omitting several elements of standard competition law where economics usually plays a key role. The paper asks where economic analysis still adds value in implementation.

2.2 Self-preferencing and the single-answer interface

At the mouth of the river, a second strand asks not how a gatekeeper treats rivals seeking access, but how an intermediary treats them once inside. Heidhues, Köster and Kőszegi (2023) sharpen the welfare assessment, where earlier work assumed rational consumers and concluded that steering, holding prices fixed benefits them, they show that for fallible consumers, who

make statistical and strategic mistakes in evaluating offers, steering based on high-quality information about those mistakes rather than about their preferences is typically harmful, and sometimes drastically so. Applied to platforms, this is self-preferencing, and it explains why the value of holding a user's AI agent exceeds what advertising alone would suggest: an AI agent that is advising the user can shape what the user reaches, and a gatekeeper integrated downstream collects the difference.

In recent years, a shift in the nature of digital outputs has become increasingly apparent. Search engines, traditionally serving as gateways to information, are evolving into "answer engines" that leverage integrated LLMs to provide a single, prominent response directly within the search interface. The transition from a set of competing results to a singular AI-generated answer raises concerns regarding user choice and market access. Similarly, AI agents are emerging as widely used single-result interfaces. In this context, practices such as self-preferencing and limited interoperability may reinforce the position of gatekeepers in ways that were not fully anticipated by the Digital Markets Act (DMA). Peitz (2022) provides a meaningful reading of the Article 6(5) prohibition: self-preferencing, showing that it reaches favourable treatment in ranking, indexing, and crawling, and decisively for AI, that the statutory definition of ranking already addresses virtual assistants and the case where only a single result is returned. He distinguishes three configurations the ban must handle (favouring one's own products to end users, favouring one's own ancillary service such as a payment system to business users, and indirect self-preferencing where using the gatekeeper's input improves a third party's ranking). The article also highlights one central, unresolved difficulty: separating illegitimate bias from legitimate, quality-based differential treatment, which is hardest precisely when ranking is generated by self-learning systems. Carugati (2022) pins down why the "fair, transparent, non-discriminatory" standard is so hard to apply, and proposes a testable parameter-list approach, arguing that because the DMA defines neither equal treatment nor self-preferencing, and because the Commission cannot observe the gatekeeper's reasons, authorities should proceed with a list of objective parameters tested independently. With AI-driven ranking/recommendation, the parameters become even more opaque, making the "objective and unbiased" test even harder.

Hovenkamp (2022) in counterpart argues that mild self-preferencing is ubiquitous and not inherently anticompetitive, that surfacing one's own product captures only consumers who are indifferent rather than those who strictly prefer a rival, and that ranking quality is irreducibly subjective. He defends the app store walled garden through the logic that: devices are sold cheaply and profits recovered from intensive service users through transaction fees, for which a closed system is needed to collect. And treats a statutory ban on platform vertical integration as deeply misguided because integration yields efficiencies and internal expansion is itself a form of entry. This review takes issue on three fronts. First, and most importantly, the magnitude point: a point his own concession exposes, since he accepts that demoting a rival to a results page no one reaches is functionally exclusion. An interface that returns a single answer maximises exactly that magnitude. Second, the vertical-integration claim: Hovenkamp treats

the efficiency case as decisive and dismisses the conflict-of-interest concern, but in doing so he ignores the gatekeeper's dual-role incentive to foreclose that Bourreau (2022) and Scott Morton et al. (2023) describe and that Padilla, Perkins, and Piccolo (2022) model formally; indeed his claim that losing the service monopoly would merely raise device prices is directly contradicted by that model. Third, his metering defence assumes the walled garden is the only mechanism by which a platform can meter and monetise, which is debatable on its own terms. Indeed, in Apple's case, the revenue from hardware is significantly higher than that from services: in fiscal 2025, Products net sales of roughly $307 billion were about 3 times Services net sales of roughly $109 billion (Apple Inc., 2025).

Recent works move the debate onto AI assistants specifically: Morozovaite (2023) shows how voice assistants can convert advanced behavioural influencing "hypernudging" into a subtle, largely invisible form of self-preferencing, extending the Article 102 analysis to a setting where the choice is effectively made for the user. The framing apparatus comes from the DMA itself, which delineates core services by purpose rather than device and so treats the App Store and Safari as unified across devices while keeping the operating systems as distinct services, and from Fletcher et al.'s (2024) "different purposes" test, the tool for deciding whether an AI agent differs enough from an answering assistant to give them separate treatment. We have to keep in mind that most of the DMA was initially written for assistants that answer rather than agents that act. This whole debate concerns what an agent does once it holds the user. How it comes to hold the user is settled upstream.

2.3 Ecosystem levers: distribution, data, and multihoming

A second body of work concerns the structural levers by which a mobile gatekeeper could entrench an AI agent rather than solely bias its outputs. The first lever is distribution: pre-installing and defaulting to the gatekeeper's own agent and tying it to the device. It is the route Bostoen and Krämer (2026) judge most pressing, and they locate the chokepoint at the operating-system layer, already addressed by the uninstallation and default-choice obligations of Articles 6(3) and 5(7) and by the interoperability duty of Article 6(7).

The second lever is data, Hagiu and Wright (2025) make the case for optimism: data feedback loops are weaker than commonly assumed, because much training relies on public or purchasable data that generates no loop, data is also often non-unique and substitutable, and learning curves are thought to plateau. They distinguish across-user learning, which can generate data network effects and tip a market, from within-user learning, which produces lock-in but not tipping, and conclude that foundation models will sustain several well-funded competitors rather than a single winner. On agents they are double-edged: an agent that transacts for the consumer could be beneficial for its user by cutting the cost of multihoming, therefore intensifying competition. But a powerful agent provider for instance a gatekeeper might self-preference its other services or products.

Bostoen and Krämer (2026) first separate an agent from an assistant along several dimensions: autonomy, learning from experience through reinforcement learning, persistent memory, and the capacity to perceive and act through tools. They argue that, for such systems, data-driven network effects do hold after all: reinforcement learning needs user feedback, conversational data is a competitive asset, and a personal agent must learn idiosyncratic preferences. The two papers reach opposite conclusions, and that is the hinge of the present argument: the optimistic reading holds for foundation models, but agentic AI may flip the result. Bostoen and Krämer then map the agent onto the technology stack: chips, operating system, cloud, model, and agentic framework across which vertical integration is likely, and identify two harms tracking the DMA's twin goals: foreclosure of an agent and foreclosure through an agent (self-preferencing). Their headline conclusion is that the DMA is "surprisingly future-proof,". Their conclusion that the existing toolkit reaches agents holds (as the legal analysis below shows, recent enforcement confirms it does), we build from their conclusion on two grounds drawn from the sources themselves. First, the optimism is in tension with the paper's own analysis: the authors concede that vertical integration across the stack makes concentration likely and that the virtual-assistant fit degrades as autonomy rises, while acknowledging that the category's applicability owes more to legislative coincidence than to design. Second, the premise that "the operating system is already designated" is more fragile than it appears. The Commission treats the operating system of each device differently, so the operating-system chokepoint is, in practice, more fragmented than the paper implies, a fragmentation that bears directly on the mobile-segment. That fragmentation is not merely formal: because these platforms carry obligations of differing weight, it is precisely what allowed Apple to release Siri AI on its less heavily regulated surfaces (macOS and visionOS) while withholding it from the more heavily regulated iOS and iPadOS. Whether this is, on balance, a desirable consequence is genuinely debatable. On an optimistic reading, the fragmentation lets regulation bite where the competition concern is gravest, while leaving lighter-touch surfaces free. On a more sceptical reading, it simply hands the gatekeeper a lever: by sorting its rollout across Core Platform Services (CPSs) of differing regulatory weight, it can deliver the integrated feature where it faces no duty to share and withhold it where it would.

Bourreau (2022) treats the interplay between multihoming and horizontal interoperability: multihoming is what lets an entrant survive against an incumbent until it reaches critical mass, yet horizontal interoperability reduces multihoming. He also points out that more interoperable functionalities should be more effective for competition but require higher cost or complexity, which could result in less differentiation, and weakened innovation incentives. Whereas on vertical interoperability, he points that gatekeepers have a dual role (OS developer and device manufacturer) that gives them the ability and the incentive to foreclose the downstream market. Scott Morton (2024) continue for app stores, multihoming and store multiplicity are a feature rather than a fragility, because niche or curated stores ride the gatekeeper's existing two-sided network effects rather than building their own, so multihoming enables differentiation rather than fragmenting the market. Opening the operating system to rival stores can therefore de-

liver both lower fees and competition in variety, but only if enforced well, because security fear can be weaponised. Padilla (2024) supplies a sharp warning: contestability obligations do not necessary mean secure fairness, because a fee structure such as Apple's fee (Developers staying exclusively in the App Store keep the old 15-30% fees on app and in-app transaction; those who multi-home in alternative stores face a new fee including a per-install fee = de facto minimum fee per user) can be formally compliant while rendering rival entry non-viable. Scott Morton et al. (2023) generalise the remedy these papers converge on. They cast equitable interoperability as the regulatory "supertool" against network effects, built around the distinction between competition for the market and competition in the market. They then extend it explicitly to the partial interoperability that voice and IoT assistants grant third-party devices, while leaving open whether the shared interface should be provided free of charge. What they concede matters here: equitable terms are hard to police, because discrimination can be hidden in the implementation. That concession sets the terms of the last stretch of the river. Everything reviewed so far assumes that whatever access a rival is granted can, in principle, be seen and assessed. Push the question far enough upstream and the assumption fails: the implementation in which discrimination can be buried is, at the source, the runtime and the scheduler of the device itself.

2.4 The hardware layer

To grasp what is at stake in the NPU, and its importance in the provision of on-device AI, we must turn to the engineering literature, which has studied it closely. Its preoccupations are well settled: the resource constraints of devices that carry only a few, often workload-specific, accelerators; the model-compression techniques that make inference feasible on such hardware; the energy and sustainability cost of running models at scale; and the privacy and security implications of processing sensitive data locally. Its consumer application: the on-device assistant (Singh & Gill, 2023), the agent layer on which the whole competitive question turns. And, strikingly, this literature does not confine itself to engineering. Meuser et al. (2024) devote a dedicated stakeholder analysis to the role of government, observing that European actors in particular will press privacy and security regulation the most.

The unfollowed step from the economics of on-device inference to the question of who controls the hardware on which it runs, and on what terms a rival may use it is the part this thesis focuses on: the interoperability of on-device AI and the hardware it needs to run, the highest level of integration. As inference migrates from the cloud onto the mobile device, an agent's functioning depends on dedicated hardware: the CPU, the GPU, RAM, and above all the AI accelerator, or NPU. Whether the NPU is Apple's Neural Engine, Google's Tensor, or Qualcomm's Hexagon, a firm that controls both the operating system and the accelerator can in principle give an advantage to its own agent.

Recent enforcement supplies both an analogy and an early warning. In its Article 6(7) spe-

cification proceedings the European Commission set out how Apple must provide interoperability with hardware and software features of iOS but the request-based process and the integrity exception need to ensure that access is in practice denied on technical grounds, raising Bourreau's (2022) pretext warning. The engineering literature is equally explicit about the component on which this migration depends. Lee (2021) traces how deep neural networks, once accelerated chiefly by GPU-equipped servers, came to require dedicated, energy-efficient hardware as their applications moved down to mobile nodes, and how the NPU emerged as the hardware purposely built for that role. The technical basis for the division is precise: training a model demands high numerical precision, whereas inference tolerates much lower bit-precision, which is exactly what allows it to run at low power on a dedicated accelerator rather than a general-purpose processor (Lee, 2021). This literature therefore establishes the NPU not as another processing unit within a system-on-chip, but as the hardware component engineered to support efficient on-device AI inference.

2.5 From the literature to the question

Taking the mobile device market as its focal case, the thesis asks whether the DMA effectively reaches a gatekeeper's privileged access to its hardware, and especially to its NPU, for its own agent, or whether the obligation's access-not-parity scope, its integrity exception, and its request-based enforcement leave a hardware-layer foreclosure gap. This is a forward-looking theory of harm, built by analogy from the current European Commission proceedings; but, given the DMA's own anticipatory design, it is exactly the kind of question the instrument was meant to be tested against, and the place where the literature's thinning waters give out is the place this thesis begins to fish.

3 Technical grounds

The argument that follows turns on a claim about hardware, and that claim cannot be assessed without a shared vocabulary. This chapter supplies it, and fixes three things the rest of the thesis relies on: what an AI agent is, and how it differs from the assistant the DMA was drafted around (Section 3.1); which on-device resources its performance actually depends on, and through which interfaces they are reached (Section 3.2); and the distinction between reaching a resource and reaching it on equivalent terms (Section 3.3). The last of these is the pivot. It is what the access-quality index θ of Chapter 4 measures, and what Article 6(7) is read against in Chapter 5.

3.1 The AI layer

3.1.1 Inference vs training

Inference and training are the two main operations behind a complete AI model: fundamentally different, yet complementary. Training is the one-off, upfront process of building the model: enormous datasets are pushed through the network until its parameters perform. It is compute-intensive, runs for long periods of time on clustered GPUs in data centres and is done by the firm owning the model. Inference is what happens every time the finished model is used: a query arrives, the model runs a single forward pass, and returns an output.

The two pull toward different hardware and different places. Training rewards raw parallel throughput and tolerates latency, cost and power, so it lives on cloud GPUs. Inference for an interactive on-device agent has the opposite priorities: low latency, low cost at scale, privacy, offline availability and battery efficiency which is exactly what an NPU is built for. Although inference can be performed on dedicated cloud servers, it increasingly migrates onto the device and onto the NPU, while training remains in the data centres.

3.1.2 Agentic-AI

Although the term "agentic AI" has become increasingly common in recent literature, it is, as Bostoen and Krämer (2026) note, not yet well-defined. Because it is used intensively throughout this thesis, we set out a working definition here, building on theirs.

Three notions form a ladder of increasing capability, and four features mark the steps that lead to the agent: autonomy, persistent memory, learning from data, and the use of tools.

Autonomy. A system with minimal autonomy executes only what has been explicitly requested and involves human interaction for deeper requests or adjustments. A system with high autonomy, by contrast, interprets what has been requested and acts on the user's behalf, making its own decisions about what to do and how while involving minimal interaction with its user. On a spectrum of autonomy, bare generative AI generally sits at the bottom, an assistant sits low to intermediate depending on its degree of integration, and an agent ranges considerably higher.

Persistent memory. A second feature is the ability to store information across interactions: the topics of previous conversations, data gathered in earlier interactions, or simply how the user prefers to be answered.

Reinforcement learning and data. A further feature that separates agentic from merely generative AI is the capacity to learn through reinforcement, which requires very large quantities of data. Data is a point of tension, since a firm can withhold it from rivals at little to no cost to its own use: it can be characterised as a club good, a non-rival but excludable input. This makes it a potential chokepoint. Firms need user feedback, conversational datasets are a competitive advantage, and personal-assistant agents must learn idiosyncratic user preferences, so user-feedback data is a vital input in its own right (Bostoen & Krämer, 2026); control over such data can, in turn, shape competition at the agent layer.

Tools. Finally, the ability to use tools is among the defining aspects of agentic AI: agents perform complex tasks by mobilising the tools made available to them. For example code execution, document handling, or the use of external services and applications.

We can now distinguish three related notions: generative AI, AI assistant, and agentic AI. The three are intertwined, for instance, an agent typically relies on generative AI to produce its outputs. Yet, they are quite different, whilst they all commonly use the same engine, the large language model (LLM), an underlying neural network trained to predict text, the system built around the model, is where lies the difference:

Generative AI is the broadest umbrella: any system whose output is generated content (text, image, code, audio). It produces content by recombining patterns learned from its training data.

The AI assistant model is a generative system wrapped in a conversational interface. It waits for a prompt, answers, and stops. Today's assistants are almost always LLM-powered, but the defining feature is the interaction pattern (you ask, it responds). A voice assistant like classic Siri is an assistant; so is a firm's support chatbot.

Whereas Agentic AI / AI agent is a system that adds, on top of the model, autonomy, planning, memory, and the ability to act through tools (call Application Programming Interface (API)s, make a booking, control apps, transact). It does not merely answer prompts; it acts. Crucially, an agent is an LLM plus an additional orchestration layer that gives it the ability to perform tasks.

These distinctions above are what makes an on-device AI agent different from, say an ordinary AI agent application, which the user opens, addresses, and closes. An on-device AI agent may also occupy what is best called the *assistant slot*: the position an operating system reserves for the agent it treats as the device's own. The slot is not reached by launching an icon but by the gesture the system reserves for it, a long press on a side button, a wake word, a swipe from the corner, and it is reachable from any screen, including a locked one, without the user leaving what she is doing.

What the slot confers is not a better interface but a different class of behaviour. An agent that holds it can be summoned mid-task, from any screen and without abandoning what the user is doing, so that it is invoked in the situation itself. It can also see the situation it is being asked about, since what is on screen and what the system knows of the user are available to it as context, whereas an ordinary application must be told. It can act rather than advise: it opens the booking, sends the message, changes the setting, where an application can at best hand the user back to another application to finish the task herself. And it can do these things unprompted, waking on a condition rather than on a launch, which is what separates an assistant that is present from one that must be remembered.

The slot is also, by construction, singular and the consequence is exclusivity: the position cannot be shared, and an agent that does not hold it must be summoned deliberately, which is a materially different thing from being present. This is why the model developed in Chapter 4 treats agent choice as single-homed: the slot is occupied by whichever agent the system de-

fault names, and reassigning that default takes some nine successive steps through the system settings[2] rather than being a choice made in the moment, so it is made once, if at all, and then forgotten.

The competitive stake follows from what the occupant does. An agent in this position mediates the user's access to the market: which application opens, which service is called, which product is surfaced, which transaction is completed. Holding the slot is therefore worth more than the usage it directly generates, and it is that surplus, rather than the agent's own subscription revenue, that the model's monetisation parameter is meant to capture.

3.1.3 On-device AI

On-device, or edge, AI refers to running artificial-intelligence computations on the local device rather than on a centralised cloud server (Singh & Gill, 2023). It spans a wide range of hardware including sensors, industrial machinery, autonomous vehicles, and mobile devices. The shift from cloud to device gives some advantages: lower latency, the round trip to a remote server is eliminated and results are returned in milliseconds; data privacy, sensitive inputs do not need to leave the device; offline availability, a model that does not depend on the network can operate without a connection; and reduced bandwidth and transmission cost (Singh & Gill, 2023; Meuser et al., 2024).

Running a modern model on a device is not trivial, making inference feasible therefore relies on two things. The first is a set of techniques that shrink the model to fit the device, model compression methods such as pruning and quantization, which reduce a model's size and numerical precision with limited loss of accuracy. The second is a dedicated on-device runtime that schedules the work onto whatever hardware is available, such as Apple's Core ML or the TensorFlow Lite lineage on Android (Singh & Gill, 2023). This combination is what allows inference to move onto the mobile device at all and it is at this second layer that the gatekeeper can shape how well a rival's agent performs. This migration carries a competitive significance, for a platform owner that both controls the device and offers its own agent, this reshapes the economics of access: the accelerator is a scarce, shared resource that the operating system must arbitrate, and the firm doing the arbitrating decides how much of it a rival's agent may use. The accelerator's role is considerable, and so is the incentive to deny rivals access to it.

3.2 The chip vocabulary

3.2.1 NPU (Neural Processing Unit) / AI accelerator

An NPU is a specialised processor designed to accelerate the execution of deep neural networks. Unlike general-purpose processors, NPUs are specifically optimised for the low-precision multiply-accumulate (MAC) operations that dominate neural network inference.

[2]Observed on Android 15 in July 2026; the exact path varies by manufacturer and version.

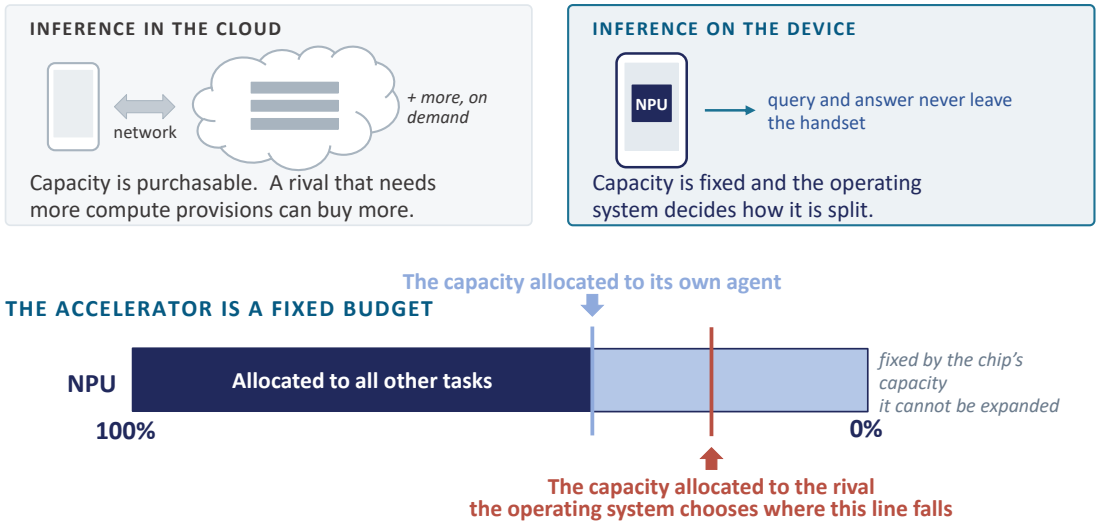

Figure 2: Cloud and on-device inference. Where inference runs in the cloud, capacity is purchasable and therefore extendable. Where it runs on the handset, capacity is fixed by the chip and cannot be expanded, so the NPU must be divided between the tasks competing for it. The migration of inference onto the device therefore turns a purchasable input into an allocated one, and the firm that arbitrates the allocation competes in the market being allocated.

Their architecture combines massive parallelism with specialised processing elements, efficient memory hierarchies and optimised dataflow strategies to maximise data reuse while minimising costly memory accesses. As memory movement often consumes more energy than computation itself, modern NPUs prioritise reducing data transfers through on-chip memory and tailored dataflow architectures. They also rely on reduced numerical precision, which significantly improves performance and energy efficiency while maintaining high inference accuracy. These architectural characteristics enable NPUs to execute deep learning workloads with substantially lower power consumption and higher throughput than conventional CPUs or GPUs, making them particularly well suited for on-device AI applications (Lee, 2021).

Driven by the rapid expansion of artificial intelligence, Neural Processing Units (NPUs) have become a standard component of modern system-on-chips (SoCs) across smartphones, tablets, PCs and other connected devices. As mentioned above, unlike CPUs or GPUs, NPUs are enabling complex AI models to run efficiently directly on the device, this hardware capability is central to this thesis. Indeed, since the NPU is not an independently accessible resource, third-party applications cannot interact directly with the hardware but must instead rely on software frameworks provided by the operating system, such as Apple's Core ML or Android's Neural Networks API (NNAPI). As a result, the operating system effectively controls access to the NPU by determining which functionalities are exposed to rivals and under what conditions.

The NPU, despite having a central role in the inference of on-device AI is nevertheless one

of several hardware resources that determine the performance of on-device AI. This broader perspective is reflected in the European Commission's Google proceeding (Case DMA.100220), where the Commission identifies access to "hardware resources such as the CPU, GPU, NPU and RAM," together with background execution capabilities, as relevant technical resources that may affect the ability of third parties to compete in AI-enabled services.

3.2.2 The other key components: RAM, RAM residency, background-execution, CPU and GPU

RAM (random-access memory) is the device's short-term working memory, where a running model and the data it is processing are held. For an AI model the binding constraint is how much it is allocated and, crucially, whether it may remain resident; data kept loaded in memory between uses rather than lost and reloaded. A model that stays resident responds instantly; one that must be reloaded on each invocation is slower and consumes more energy. RAM residency is therefore a quality-of-service lever in its own right, and it is precisely the one the Commission singled out in the Google proceeding.

Background execution is the ability of a process to run when it is not the app in the foreground, to act without the user having the app open. An agent that may run in the background can be ambient and proactive; one confined to the foreground can act only when explicitly opened, which for an agent is a severe handicap.

The CPU is the device's general-purpose processor acting as the brain of the device and the GPU is hundreds or thousands of smaller, less complex cores, it is optimised for high-throughput parallel processing. Both can run inference when the NPU is unavailable or not exposed. Their relevance here is as the fall-back path: a rival agent denied efficient NPU access is pushed onto the CPU or GPU, where the same task runs more slowly and drains more battery.

Taken together, these resources are a bundle of scarce on-device capabilities each of which the operating-system owner allocates, and on each of which it can favour its own agent. This is why the gatekeeper does not need to close a single door to foreclose a rival; it can simply allocate less generously to others than to itself on one of the component, the NPU being the central one regarding on-device AI computation.

3.3 Access and parity

That bundle forces a distinction the rest of the thesis turns on, and which the wording of Article 6(7) does not make on its face.

Access to a good or service is whether a rival can reach it at all. It is binary: either the rival has access or it does not. Under Article 6(7) DMA, outright access cannot be denied. Parity, by contrast, is whether the rival reaches the service on quality-equivalent terms. Foreclosure at this layer is therefore mostly a matter of degraded access rather than outright denial. The levers are not a binary on/off; they are:

- Non-exposure → never publishing the API the first party uses. This is the closest to outright denial.
- Capability gating → exposing a reduced API: fewer operations, smaller models, lower precision, no access to the newest features the first-party model uses.
- Scheduling / priority → the OS scheduler decides whose workload runs on the NPU and when; a third-party job can be deprioritised, in which case it falls back to the slower CPU or GPU path.
- Entitlements → high-value capabilities are often gated behind the gatekeeper's will. Third parties must request access and the gatekeeper can refuse it; background execution and always-on use.

Three things follow from this chapter, and the rest of the thesis rests on them. First, what is contested on the device is inference rather than training. Second, the firm doing the allocation is the operating-system owner, since the accelerator is reachable only through interfaces that firm publishes, and it competes in the very market it arbitrates. Third, the position worth allocating to is singular. The assistant slot cannot be shared, which is why the next chapter treats agent choice as single-homed. Chapter 4 collapses theses levers into a single index, θ, and asks what a profit-maximising gatekeeper does with it once outright denial has been removed from its choice set.

4 The baseline model

4.1 Purpose and positioning

This chapter develops a simple formal model of the thesis's central claim: that an interoperability obligation which secures *access* to on-device AI accelerators without securing *parity* leaves the gatekeeper an exclusionary instrument: the quality of that access, and that the incentive to use it is governed by the forces identified in the literature reviewed above. The market has an explicitly *vertical* structure. Upstream sits a device manufacturer, a monopolist over the hardware, who controls the terms on which anything reaches the accelerator; downstream sits the market for *AI agents* that run on that hardware, in which the manufacturer's own AI agent competes with a rival's. This architecture is the same as the legal architecture of the thesis. The upstream link is the object of Article 6(7) (vertical interoperability: the terms on which the rival reaches the hardware), and the downstream market is where Article 6(5) operates (self-preferencing and steering in the AI agent's output).

Two modelling choices sharpen the exercise relative to a generic duopoly. First, the device price is *endogenous*, and the manufacturer is a monopolist who extracts the consumer surplus generated on the device through that price. Devices and AI agents are complements: a device is

worth more when the AI-agent layer on it is better, and the monopolist prices that complementarity in. Second, the two AI agents are normalised to identical intrinsic quality ($q_G = q_R = 1$). Nothing distinguishes them except θ, the access the gatekeeper grants: whatever asymmetry the market displays is *manufactured* by the gatekeeper's choice, not inherited from product differences. Two boundaries are kept from the earlier discussion. Assumption 1 combines an interior constraint, which keeps the rival's share positive at every admissible access quality, and a covered-market condition, which ensures the consumer least well served still prefers one agent to none.

4.2 Setup

Vertical structure and players. An upstream *device manufacturer* (the gatekeeper, G) is a monopolist over the device and its operating system. It sells the device to consumers at a price p of its choosing. Downstream, two *AI agents* can run on the device, but only one at a time: the manufacturer's own, A_G, and a rival developer's, A_R. AI agents are supplied to consumers at a price of zero, as in practice; each firm monetises usage of its own AI agent at a rate $\alpha > 0$ per user.

The quality-lowering instrument. The rival's AI agent reaches the device's NPU only through interfaces the manufacturer's operating system controls. The manufacturer chooses the rival's *access quality* $\theta \in [\theta_{\min}, 1]$, where $\theta_{\min}$ is the regulatory floor. The two AI agents have identical intrinsic quality, normalised to one, so delivered qualities are

$$v_G = 1, \qquad v_R = 1 * \theta.$$

The only source of asymmetry in the downstream market is the gatekeeper's access choice; its own AI agent always runs at full quality, and lowering θ costs it nothing. The index θ is not an abstract quality parameter: it collects the parameters along which an operating system determines how well a third-party AI agent performs, each of them identified in the technical chapter above and each of them the object of the Commission's proposed measures in Case DMA.100220.

Consumers. There is a unit mass of consumers. Each values the device itself at $v > 0$, identical across consumers. Tastes over AI agents are horizontally differentiated à la Hotelling: a consumer's location x is drawn uniformly on $[0,1]$, with A_G at 0, A_R at 1, and a transportation (mismatch) cost $t > 0$ per unit of distance. A device owner at x obtains agent-layer utility $1 - tx$ from using A_G, $\theta - t(1-x)$ from using A_R, and 0 from using neither.

Timing. (1) The regulator sets $\theta_{\min}$; the manufacturer publicly chooses θ and the device price p. (2) Consumers decide whether to buy the device. At this stage they are *ex ante identical*: a

consumer learns her taste location x only after acquiring the device,[3] so her willingness to pay is v plus the *expected* agent-layer surplus, $v + W(\theta)$, which she computes with correct foresight of stage 3. (3) Each device owner learns x and uses exactly one AI agent: the device offers a single assistant slot, since invoking a different AI agent requires changing the system default, so agent choice is single-homed.

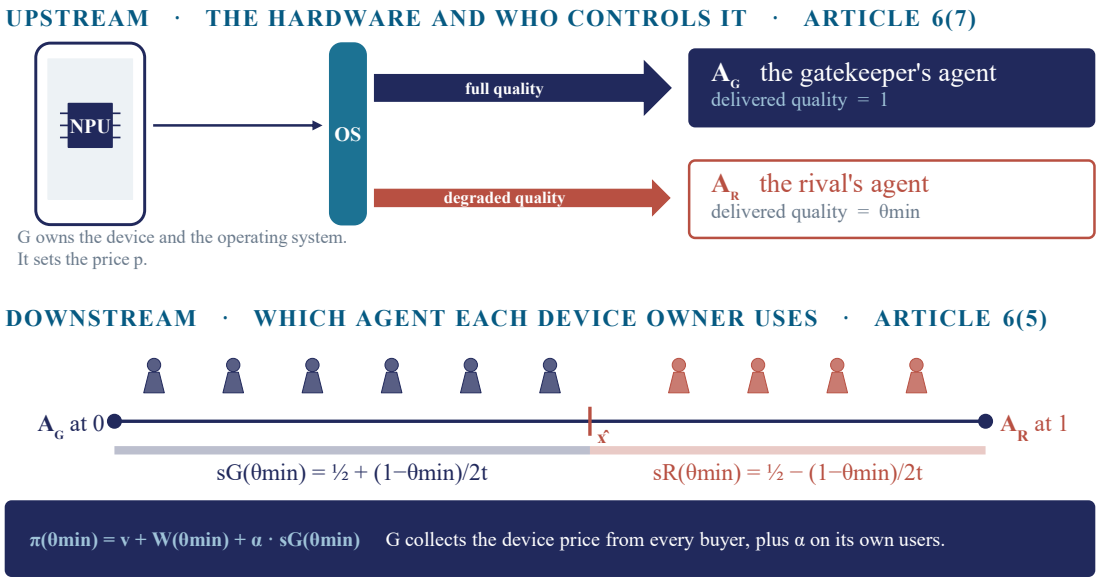

Figure 3: Structure of the model, drawn at the floor-sitting equilibrium $\theta = \theta_{\min}$. The accelerator is reachable only through the operating system, which the gatekeeper also owns. From that single chip two channels lead to two agents of identical intrinsic quality, and the only difference between them is the access quality θ the gatekeeper grants the rival. Downstream, each device owner uses one agent, so the split of the unit mass of consumers at $\hat{x}$ is the split of the market.

Solution concept. The game is solved by backward induction, and the equilibrium reported throughout is its subgame-perfect Nash equilibrium. Stage 3 determines, for any θ, which AI agent each device owner uses, and hence the shares $s_G(\theta)$, $s_R(\theta)$ and the agent-layer surplus $W(\theta)$ (Lemma 1). Stage 2 determines, for any pair (θ, p), whether consumers buy, given the $W(\theta)$ they correctly anticipate from stage 3. Stage 1 then selects (θ, p) to maximise the manufacturer's profit, taking the stage-2 and stage-3 responses as given. Consumers hold correct expectations at every stage, so no player's continuation behaviour rests on a threat that would not be carried out.

[3] Buyers do not know in advance which AI agent will suit them: a device is chosen once and held for years, whereas the fit between a user and an agent is discovered through use, over a landscape of agents that turns over far faster than the device replacement cycle. What consumers price at purchase is therefore the expected quality of the agent layer the device gives access to, not the agent they will end up using.

Device pricing: extracting the complementarity. The device price is the solution to the manufacturer's stage-1 problem, given the stage-2 behaviour it anticipates. Normalise the marginal cost of the device to zero. At stage 2, consumers are identical and each buys if and only if her participation constraint holds,

$$v + W(\theta) - p \;\geq\; 0,$$

so device demand is the step function $D(p) = 1$ for $p \leq v + W(\theta)$ and $D(p) = 0$ otherwise. Anticipating this, the manufacturer solves

$$\max_{p} \Big[p + \alpha\, s_G(\theta)\Big] D(p).$$

The manufacturer's profit is $p + \alpha\, s_G(\theta)$ as long as the consumer buys, and zero otherwise: a consumer who does not buy the device runs no AI agent on it, so the downstream revenue disappears with the upstream sale. Below $v + W(\theta)$ the profit is strictly increasing in p, since raising the price does not cost a sale; above it, demand is zero and so is profit. The optimum is therefore the highest price the consumer will accept, at which the participation constraint binds:

$$p^*(\theta) \;=\; v + W(\theta). \tag{1}$$

Serving the market dominates not serving it, since $v + W(\theta) + \alpha\, s_G(\theta) > 0$ under Assumption 1. What it delivers is the complementarity between the device and the AI agents, priced in: a better AI-agent layer raises W, and the monopolist sells that improvement back to consumers.

Payoffs. Let $s_G(\theta)$ and $s_R(\theta) = 1 - s_G(\theta)$ denote the AI agents' usage shares among device owners, and $W(\theta)$ the expected agent-layer consumer surplus. The manufacturer earns the device price upstream and the AI-agent monetisation on its own users downstream:

$$\pi(\theta) \;=\; \underbrace{p^*(\theta)}_{\text{upstream}} + \underbrace{\alpha\, s_G(\theta)}_{\text{downstream}} \;=\; v + W(\theta) + \alpha\, s_G(\theta), \tag{2}$$

while the rival developer earns $\alpha\, s_R(\theta)$. The two terms of (2) are the thesis's two channels: $\alpha\, s_G$ is the *AI-agent monetisation* (business-stealing) channel, and $W(\theta)$ is the *hardware* channel a better AI agent offer raises the surplus the device generates, which the monopolist harvests through p^*.

Assumption 1 (Interior shares and covered AI-agent market)**.**

$$1 - \theta_{\min} \;<\; t \;\leq\; 1 + \theta_{\min}.$$

The first inequality guarantees that both AI agents retain positive usage share at every feasible θ; the second guarantees that the consumer indifferent between the two AI agents still

prefers using one to using none. Assumption 1 is also what separates this chapter from refusal-to-supply models such as Fumagalli and Motta (2020): total exclusion is not in the gatekeeper's choice set, as the Regulation intends; what remains in its choice set is how badly the rival's AI agent runs.

4.3 The static stage: access is not parity

Lemma 1 (Shares and surplus, tracking transportation costs)**.** *Under Assumption 1, the consumer indifferent between the two AI agents is located at*

$$\hat{x} \;=\; s_G(\theta) \;=\; \tfrac{1}{2} + \frac{1-\theta}{2t}, \qquad s_R(\theta) \;=\; \tfrac{1}{2} - \frac{1-\theta}{2t},$$

and the expected AI-agent-layer consumer surplus, net of transportation costs, is

$$W(\theta) \;=\; s_G(\theta)\cdot 1 + s_R(\theta)\cdot\theta \;-\; \frac{t}{2}s_G(\theta)^2 \;-\; \frac{t}{2}s_R(\theta)^2 \;=\; \frac{1+\theta}{2} - \frac{t}{4} + \frac{(1-\theta)^2}{4t}.$$

Moreover,

$$\frac{ds_G}{d\theta} = -\frac{1}{2t} < 0 \qquad \textit{and} \qquad \frac{dW}{d\theta} = s_R(\theta) > 0.$$

Proof. The indifferent consumer solves $1 - t\hat{x} = \theta - t(1-\hat{x})$, giving $\hat{x} = \frac{1}{2} + (1-\theta)/2t$. Aggregating each consumer's delivered quality net of her transportation cost,

$$W(\theta) = \int_0^{\hat{x}} (1-tx)\,dx + \int_{\hat{x}}^1 \big(\theta - t(1-x)\big)\,dx = s_G + \theta\, s_R - \frac{t}{2}\hat{x}^2 - \frac{t}{2}(1-\hat{x})^2,$$

and substituting $\hat{x} = s_G(\theta)$ yields the closed form. Differentiating in θ: the terms in $d\hat{x}/d\theta$ cancel by the indifference condition at $\hat{x}$ (an envelope argument relocating the exactly indifferent consumer creates no first-order surplus change), leaving only the direct effect on the rival's users, $dW/d\theta = s_R(\theta)$. □

Raising the rival's access quality *shrinks* the gatekeeper's AI-agent share, and *raises* the surplus generated on the device. These opposing movements are exactly the tension in (2). Substituting,

$$\frac{d\pi}{d\theta} \;=\; s_R(\theta) - \frac{\alpha}{2t}, \qquad \frac{d^2\pi}{d\theta^2} \;=\; \frac{1}{2t} \;>\; 0. \tag{3}$$

Lemma 2 (Convexity and corner choice)**.** $\pi(\theta)$ *is strictly convex on* $[\theta_{\min}, 1]$*; the gatekeeper's optimum is a corner,* $\theta^* \in \{\theta_{\min}, 1\}$.

The marginal cost of openness (AI-agent users lost to the rival, worth $\alpha/2t$ per unit of θ) is constant, while the marginal benefit (the rise in device-extractable surplus, $s_R(\theta)$) is itself increasing in openness: the more users the rival's AI agent serves, the more surplus each further improvement in its delivered quality creates, all of it harvested through the price. Openness

therefore has a tipping logic and *interior* degradation, when observed, is supplied by the regulation rather than chosen freely.

Because $d\pi/d\theta$ is linear in θ, the corner comparison is exact and the threshold has a closed form.

Proposition 1 (Floor-sitting: an access standard fixes the level of degradation)**.** *Define*

$$\alpha^{\dagger} \equiv t - \frac{1-\theta_{\min}}{2}, \qquad \textit{so that} \qquad \pi(1) - \pi(\theta_{\min}) = \frac{(1-\theta_{\min})\left(\alpha^{\dagger} - \alpha\right)}{2t}.$$

If $\alpha > \alpha^{\dagger}$ — AI-agent monetisation dominates the device channel — the gatekeeper sets $\theta^* = \theta_{\min}$: access is granted at exactly the regulatory floor, and the legal definition of "effective" interoperability becomes the market outcome. If $\alpha \leq \alpha^{\dagger}$, it voluntarily grants full parity, $\theta^* = 1$.

The force of the result comes from what has been stacked against it. The gatekeeper is a monopolist that extracts the *entire* surplus of the device ecosystem through the price: it fully internalises every improvement in the rival's AI agent offer. Degradation still occurs whenever $\alpha > \alpha^{\dagger}$, because the business-stealing gain increases privately while part of what degradation destroys is the *rival's* profit, which the gatekeeper does not internalize.

Corollary 1 (What moves the threshold)**.** $\alpha^{\dagger} = t - (1-\theta_{\min})/2$ *is increasing in $\theta_{\min}$ and in t. (i) Raising the access floor not only mechanically raises the rival's delivered quality; it raises $\alpha^{\dagger}$ and can therefore* tip *the gatekeeper from floor-sitting into voluntary full parity. A stricter standard is self-reinforcing. (ii) Stronger horizontal differentiation (higher t) makes business stealing harder and voluntary parity likelier; as AI agents become closer substitutes, the foreclosure region expands.*

4.4 The steering downstream

Steering as part of α: the second foreclosure. The per-user monetisation α[4], for a vertically integrated gatekeeper, is not limited to advertising or subscriptions. The AI agent that wins a user intermediates that user's access to the market. A gatekeeper integrated downstream can *steer* those choices toward its own products and services, and the rents this steering generates are part of what a user of its AI agent is worth. Two consequences follow directly from Proposition 1. First, since floor-sitting occurs precisely when $\alpha > \alpha^{\dagger}$, steering rents (by raising α) enlarge the foreclosure region: the gatekeepers most integrated downstream, for whom the steering component is largest, are the most likely to degrade rivals' access. Second, the harm now has two distinct sets of victims: rival AI agents, foreclosed at the access layer by θ; and third-party sellers, foreclosed at the downstream layer by the steering itself.

[4]It is worth noting here that α is plausibly large for the firms concerned: these are gatekeepers whose core competence is converting user interactions into revenue at scale, through advertising and intermediated transactions. Alphabet, for instance, derives the overwhelming share of its revenue from advertising alone.

The legal mapping runs along the same vertical diagram the model is drawn on. The access quality θ, upstream, is the object of Article 6(7); the steering, downstream, is the object of Article 6(5). Reading the steering rent as a component of α shows why the two provisions are complements rather than substitutes. Enforcing Article 6(5), constraining self-preferencing in the AI agent's output caps the steering component of α, lowers the gatekeeper's effective per-user monetisation, and can therefore tip it below $\alpha^\dagger$ into voluntary parity: policing the downstream symptom weakens the upstream incentive for the disease. Conversely, enforcing Article 6(7), raising $\theta_{\min}$, both raises the threshold $\alpha^\dagger$ (Corollary 1) and shrinks the captive user base whose choices a biased AI agent can exploit, reducing what steering is worth. Each provision, enforced, weakens the conduct the other polices; enforcing neither lets the two compound.

Three results carry into the chapters that follow. First, the gatekeeper's choice of access quality has no interior optimum. It sits at one of two corners: full parity when the surplus a better rival agent generates on the device outweighs the per-user monetisation, and exactly the regulatory floor when per-user monetisation dominates instead. Second, raising that floor is self-reinforcing, since it lifts the very threshold above which floor-sitting is profitable (Corollary 1). Third, with α and the steering downstream, the upstream duty of Article 6(7) and the downstream prohibition of Article 6(5) act as complements rather than substitutes. All three are predictions about conduct rather than statements. The next chapter looks at how these provisions have played out in practice, in the Commission's Android proceeding and in Apple's withdrawal case.

Notation. G: the gatekeeper (upstream device manufacturer); A_G, A_R: the gatekeeper's and the rival's AI agents; $q_G = q_R = 1$: normalised intrinsic qualities of the two AI agents; θ: rival's access quality to the NPU; $\theta_{\min}$: regulatory floor on access quality; $v_G = 1$, $v_R = \theta$: delivered qualities; v: intrinsic device value; p: device price, $p^*(\theta) = v + W(\theta)$ in equilibrium; $D(p)$: device demand; t: transportation (taste-mismatch) cost; x: consumer's taste location on $[0,1]$; $\hat{x}$: the consumer indifferent between the two AI agents; $s_G(\theta)$, $s_R(\theta)$: AI-agent usage shares; $W(\theta)$: expected AI-agent-layer consumer surplus, net of transportation costs; α: per-user AI-agent monetisation, inclusive of downstream steering rents; $\alpha^\dagger = t - (1-\theta_{\min})/2$: floor-sitting threshold; $\pi(\theta)$: the gatekeeper's profit.

5 The economics of access versus parity

The model predicts that a gatekeeper facing an access obligation either supplies access at exactly the level the obligation requires and no more, or grants full parity outright where rival access adds more to the device price than its own agent earns it. That second condition fails for the gatekeepers that monetise their own agent most heavily. Whether that prediction bites depends on what Article 6(7) actually requires of the hardware layer, which the text alone does not answer. This chapter therefore proceeds in three steps: it reads the provision against the

hardware it is supposed to reach (Section 5.1); it tests the prediction against the only current proceeding of the Commission treating interoperability in the on-device AI stack (Section 5.2); and it takes up the case in which the gatekeeper answers the obligation by withdrawing the product rather than opening the platform (Sections 5.3–5.4).

5.1 How Article 6(7) treats hardware controlled via the operating system

Everything so far treats the gatekeeper's allocation of processing capacity as an instrument. However, this graded control is not always strategic. Much of it has legitimate technical justifications: the NPU is a shared, scarce resource, and managing it requires the platform to arbitrate access. This is what makes the regulation of this market hard: the same conduct can be strategic or necessary. To counteract, Article 6(7) obliges a gatekeeper to provide, free of charge, effective interoperability to the same hardware and software features accessed or controlled via the operating system as are available to the gatekeeper's own services and hardware. The obligation is, on its face, about hardware features, and it is especially tied to features "controlled via the operating system." As the technical background established, the on-device accelerator is reached only through interfaces controlled by the operating system, so on any natural reading it is a hardware feature controlled via the operating system, and thus falls within the provision.

The practice confirms it. The Commission has already begun to enforce Article 6(7) through its specification proceedings against Apple (Cases DMA.100203 and DMA.100204), and, of particular relevance here, has opened an Article 6(7) proceeding addressed specifically to interoperability with artificial-intelligence functionality on Google's Android. On 27 January 2026 the Commission opened proceedings to specify the measures Alphabet must implement to comply with Article 6(7) "with respect to features which are relevant for third-party providers of artificial intelligence services on mobile devices," and on 27 April 2026 it adopted preliminary findings (Case DMA.100220). The Commission frames the mobile market as at a "technological inflection point, with AI services becoming central access points,"(Case DMA.100220) and concludes that interoperability is critical so that rival AI services can deliver "at least comparable performance, responsiveness, visibility, functionality, and utility" (Case DMA.100220) to the gatekeeper's own.

Apple is the strong case: on iOS the Neural Engine has historically had no public API at all, third parties reach it only indirectly, by handing a model to Core ML and letting Apple's runtime decide whether to place it on the NPU, the GPU, or the CPU. The developer does not get to choose, can not force Apple Neural Engine (ANE) placement, and can not see why a given layer ran elsewhere.

Android was more open, but the terms have recently shifted. Where NNAPI was a vendor-neutral interface that standardised the interface but left actual performance and reliability to each vendor's drivers, LiteRT now owns the runtime itself and selects accelerators automatically. It therefore relocates control: the platform now publishes the runtime, signs the drivers

and decides which chip a workload reaches. Alphabet also introduced AICore, the privileged channel through which the gatekeeper's own Gemini Nano model reaches the accelerator: the very arrangement the Commission is now examining in Case DMA.100220.

5.2 Google's proceeding: Case DMA.100220

A note on the state of the proceeding. The analysis in this chapter is based on the preliminary findings and proposed measures of 27 April 2026, which were the operative documents when this research was carried out. On 16 July 2026 the Commission adopted its final decision specifying the measures; the measures themselves have been published, but the non-confidential version of the full decision, containing the Commission's reasoning, has not yet been released. A full comparative analysis, of what the consultation changed between April and July and of the reasoning that supported it, must therefore await that publication, and is left to further research.

Case DMA.100220 is central in the understanding of the application of Article 6(7) DMA, it is the first interoperability case to reach on-device artificial intelligence, in its investigation into Alphabet's Android. Unlike the ordinary enforcement of the DMA, which sanctions non-compliance after the fact, this specification proceeding was designed to guide Google toward a decision that complies with the DMA in the first place, rather than to penalise a decision it had already made. We must also keep in mind that the proceeding is preliminary: the Commission opened it on 27 January 2026, adopted preliminary findings and proposed measures on 27 April 2026, and set implementation for 1 January 2027; the analysis therefore describes the remedy as it currently stands, not a final decision.

The decisive part of the Commission's findings, for this thesis, is the feature group concerning access to resources, because it is there that the obligation moves from access to parity. The Commission found that Alphabet's on-device model infrastructure "currently enjoys preferential access to RAM," and that Alphabet's first-party applications may benefit because their on-device-model functionality relies on "preferential access to hardware resources"; it also noted that Alphabet uses "private APIs" to access certain on-device model functionality. The proposed remedy is to require Alphabet to allocate hardware resources to its own and to third-party on-device models "according to transparent, objective, precise, and non-discriminatory rules," (§108(a)).

That finding is the regulator documenting the very mechanism the model formalises: a gatekeeper degrades a rival's access when its per-user monetisation exceeds the threshold $\alpha^{\dagger}$, and grants parity otherwise. Alphabet is the perfect example. Its revenue rests overwhelmingly on monetising user attention and user data through advertising, complemented by commissions on the transactions it intermediates; its comparative advantage is precisely the capacity to convert an interaction into revenue at scale. An AI agent sits between the user and everything she reaches on the device, a firm that monetises interactions that well has a correspondingly high α, and it is exactly such a firm that the model predicts will supply access at the regulatory floor

rather than above it. The Commission's own finding, that AICore enjoyed preferential access to RAM until it intervened, is what that prediction looks like in the field. That finding is the empirical counterpart of the theoretical result developed earlier in this thesis: absent an effective constraint, the firm that controls the accelerator supplies itself a better grade of access than it supplies others.

Yet the text alone does not secure parity. Article 6(7) is not a checklist a gatekeeper must satisfy before it may ship a device: it states a standard, and what that standard requires of memory, of compute and of background execution had to be worked out in a proceeding. So far, the parity it promises is delivered only through ex-post, case-by-case interpretation.

5.2.1 Parity as an option: the user-toggle remedy

Does the remedy, then, secure parity at the hardware layer? It does so only conditionally, and the condition is revealing. In the Commission's preliminary findings, the operative provision on hardware access does not require that a rival's model receive access equal to Alphabet's by default. It provides instead that, where Alphabet's own models are "granted preferential access to hardware resources," Alphabet "shall allow the user to remove such preferential access from Alphabet's [models] and grant it to third-party [models]" (§108(b)). The default, on the face of the provision, remains the preferential allocation the Commission found in AICore; parity is achieved only if the user affirmatively intervenes to strip the gatekeeper's advantage and reassign it.

Two features of this design deserve emphasis. First, it inverts the ordinary logic of the obligation. Interoperability under the DMA is a duty the gatekeeper owes; here the burden of producing parity is shifted onto the user, who must know that the option exists and act on it. Second, defaults govern outcomes. The overwhelming majority of users never alter advanced system settings; for them, the self-preferential default is not a starting point to be corrected but the market outcome itself. A remedy that leaves the gatekeeper favoured unless the user chooses otherwise is, for almost every device, a remedy that leaves the gatekeeper favoured.

The Commission is aware that user-mediated access can be manipulated. Its measures permit a limitation on a third party's access arising from a user action only if the user "can take the same action with the same limiting effect" against Alphabet's own models (§108(d)). But these provisions police the symmetry of user actions and the honesty of the interface; they do not disturb the prior question of what the default is before any user acts. One might defend the opt-in as respecting user choice, and note that the baseline allocation is in any event subject to the non-discrimination rule of §108(a). But, the provision itself opens with the words "Regardless of the rules referred to under paragraph (108)(a)", it explicitly contemplates that preferential access may persist notwithstanding the non-discrimination rule, which is why the opt-in exists at all. And "respecting user choice" presupposes an informed user making a deliberate choice; the reality of default behaviour is that no such selection occurs.

5.2.2 What survives specification: the carve-out, the verification problem, and the enforcement architecture

Three difficulties survive even a successful Article 6(7) decision, and together they describe the gap that remains. The first is the integrity exception. DMA.100220 does not displace it; on the contrary, the draft measures clarify, in line with Article 6(7), how the gatekeeper "may take strictly necessary and proportionate measures to ensure that interoperability does not compromise the integrity of the operating system, hardware and software features," within pre-defined timelines. The carve-out, legitimate in purpose, is preserved alongside the parity obligation and at the hardware it is hard to justify, because a competitively motivated degradation can be re-described as an integrity measure that the outside observer can hardly rebut. This is the danger Bourreau (2022) identified: an exception whose invocation cannot be verified will, at the hardware layer, predictably absorb some of the conduct the obligation was meant to prohibit.

The remedy's second vulnerability is that the standard on which it rests is, at the hardware layer, extremely difficult to verify. Non-discrimination in resource allocation is meaningful only to the extent that a departure from it can be detected; and scheduling priority, memory residency, and latency on an accelerator are, by their nature, invisible to any party outside the operating system. A rival whose model runs more slowly cannot ordinarily tell whether the system-level scheduler allocated resources on equal conditions or quietly de-prioritised it, because the allocation is not exposed.

In the passage on integrity, the measures require that Alphabet apply only conditions "with which compliance is capable of being independently verified and not exclusively within the gatekeeper's control" (§128), and that integrity measures rest on "objective and verifiable evidence" that Alphabet must retain (§130(c)). That independent-verifiability requirement is directed at the integrity conditions, the "security carve-out" and not at the core allocation obligation; there is no equivalent mechanism to verify, independently, that the non-discriminatory scheduler in fact allocates resources without discrimination. The remedy's principal verification instrument is notification: Alphabet must inform the Commission in writing, with a justification of strict necessity and proportionality, of any integrity measure it intends to take (§131). Notification and justification are forms of self-reporting; they presuppose that the gatekeeper discloses what it does, which is weaker than the ability to observe, from outside, what it does. Therefore the parity at the silicon is, under this remedy, postulated by a rule rather than confirmed by a measurement and an opt-in intended to correct an advantage that cannot be measured is a guarantee resting on a guarantee.

The same tension runs through the integrity exception, the Commission has plainly anticipated the concern, familiar from the literature, that a security rationale can serve as a pretext for exclusion. It confines integrity measures to those that are "strictly necessary and proportionate" (§127), that reflect "a genuine integrity risk" applied "in a consistent and systemic manner" (§128), and it provides that Alphabet "shall not impose a higher level of integrity on third parties

than it applies to itself" (§128), with the burden of demonstrating the risk resting on Alphabet (§130(c)). This self-consistency requirement is a real constraint: it denies the gatekeeper the ability to hold rivals to a security standard it does not accept for its own services. But the constraint operates, once again, on Alphabet's own justification and disclosure rather than on independent observation.

The third difficulty is the enforcement architecture itself. The parity now being specified on Android emerged from a reactive, feature-by-feature proceeding against a single gatekeeper, opened in January 2026, with preliminary findings in April and a binding decision in July. With measures that take effect only with Android 18 in 2027, and with Android 19 in 2028 for one of them. Parity, on this model, exists where and when the Commission has specified it one case at a time. For a market moving as quickly, as the market for on-device AI agents, an architecture that reconstructs parity case by case is structurally on the back foot; by the time equivalence is specified for one accelerator, the technical baseline will have moved, and a second gatekeeper will remain unaddressed. It is worth noting, however, that with Google's case paving the way, a faster process can be expected for those that follow.

And, for completeness, the self-preferencing prohibition in Article 6(5) which has not been mobilized in the proceeding operates downstream of all this. It prevents a gatekeeper from treating its own products more favourably in ranking and related indexing and crawling, on transparent, fair and non-discriminatory terms, and it is the provision on which the existing literature on assistant self-preferencing principally relies (Morozovaite, 2023; Bostoen & Krämer, 2026). For an AI agent it reaches the visible output (the results surfaced, the services recommended, and so on) but not the upstream hardware advantage that produced them. It catches the symptom of a biased result, not the cause of a degraded input.

5.3 Apple intelligence and Siri AI

The Google proceeding makes Apple's position easier to read, and the analysis of this case is interesting because the Apple situation differs along three dimensions. The first is procedural: there is, as yet, no equivalent Article 6(7) proceeding addressed to AI hardware resources on iOS, and whether the parity specified for Android will carry across to Apple's platform remains open. The second is technical: where Android at least exposes a public accelerator interface and an on-device-model component, Apple's Neural Engine has historically had no public API for direct access, so the baseline against which "equal conditions" would have to be specified is harder to evaluate. The third, and most striking, is conduct. At its developer conference on 8 June 2026, Apple unveiled a deeply rebuilt, agentic Siri within its Apple Intelligence system. The system is hybrid, combining on-device inference with Apple's Private Cloud Compute. Apple made it available in the European Union only on macOS and on Apple Vision Pro, while initially withholding it from iOS and iPadOS, citing regulatory constraints.

5.3.1 Siri AI and its withholding from European markets

Siri AI was the centre of Apple's WWDC 2026 announcement: an entirely rebuilt assistant, powered by Apple Intelligence, that draws on a user's messages, emails and photo library for personal context, takes actions within and across applications, and can answer questions grounded in what is currently on screen.

This brings us onto our analysis of Apple intelligence's withholding from European markets. In their press release, Apple stated that: "[U]nder EU regulators' extreme interpretation of the DMA, Apple would have to give any virtual assistant direct access to users' private data — and the ability to directly control other installed applications — as soon as Siri AI is made available in the EU, without the essential protections necessary to keep users and their data safe.". And that: "According to EU regulators, the DMA requires Apple to give any AI system nearly unlimited access to a user's device, as well as the ability to act on that access autonomously without a user's ongoing visibility and control." (Apple Inc., 2026)

Apple therefore blames the DMA for restricting its ability to deploy Siri AI in Europe, its justification being the security concerns that would arise from giving rivals access to the confidential data required to run their on-device AI on the same terms. Crucially, Apple did not simply refuse. It proposed a mechanism, a privacy-preserving intermediary it called the Trusted System Agent, intended to let rival assistants reach the same on-device capabilities as Siri AI. And it proposed to launch Siri AI in the EU while phasing in that mechanism over an eighteen-month period. The Commission rejected this proposal, along with all other proposal Apple submitted. The two sides then publicly disagreed about what had happened, and that disagreement is itself part of the evidence.

Although Apple's security concern is credible, the Commission replied. Its spokesperson, Thomas Regnier, told reporters in Brussels that the decision not to roll out Siri AI in the EU was

> *"Apple's and Apple's only"* (European Commission, 2026b),

adding that nothing in the DMA prevents Apple from introducing new products in the EU, that Apple had been unable to develop interoperability solutions meeting essential EU privacy and security standards.

The Commission's characterisation reframes Apple's "Trusted System Agent plus eighteen-month rollout" not as a constructive compliance effort but as a request for a competitive head start: a period in which Siri AI would establish itself before any rival assistant gained comparable access. The Commission made the point explicit: granting the request would mean that no AI agent other than Siri AI would have an equal chance to be chosen by iPhone users for eighteen months. The spokesperson insisted that the EU would grant no exemptions. The stakes are large: the withholding affects millions of EU iPhone and iPad users, and Apple's Europe reportable segment, which accounted for roughly 27 percent of net sales in fiscal 2025, is the second largest of the five (Apple Inc., 2025).

Two cautions on the record. First, the dispute is contested: Apple says it offered a phased technical solution that was refused; the Commission says Apple offered no compliant solution and asked to be given an exemption. Second, there is a relevant precedent that bears on whether the withholding is permanent: when Apple Intelligence first launched in the United States in late 2024, EU users were excluded at the outset and the features arrived only months later, in spring 2025, after negotiation with regulators. The current hold-back may therefore be a bargaining posture as much as a settled outcome.

Apple's principal lever in negotiating with the Commission was the security of its consumers' data. This is exactly the integrity-exception argument the legal chapter anticipated: the claim that opening accelerator- and system-level access to rival agents cannot be reconciled with user privacy. Yet the Google proceeding suggests that sharing this kind of access need not be incompatible with safety. On Android, a user can already select its default AI assistant (namely to Copilot, to Claude or to ChatGPT) and the Commission's draft measures in Case DMA.100220 require equal access to on-device resources while also preserving the gatekeeper's ability to take strictly necessary and proportionate integrity measures. Although, one qualification should be stated: the Google measures are preliminary, adopted in April 2026 and not yet implemented, so it is too early to say that interoperability there has been "proven" safe; but the current claim is that the Commission considers integrity and interoperability reconcilable through safeguards.

As the Commission's reply also suggests, Apple appears not to have invested in interfaces robust enough to let this competition occur, and to have sought principally a carve-out from the rule. That outcome is, in a sense, reassuring: the integrity exception examined in the legal chapter and given as an "option" was precisely the channel through which a gatekeeper might convert its parity obligation. Here, mobilising its strongest privacy arguments, and even proposing a structured intermediary, Apple was not granted the exemption it sought.

5.4 The "withhold rather than share" equilibrium

The outcome is that the interoperability regime bites hard enough to make Apple withhold its own integrated agent on precisely the platforms where the DMA's obligations are the heaviest iOS and iPadOS (the designated services carrying the sideloading and interoperability duties) without thereby delivering equivalent access to rivals. The result is a "withhold rather than share" equilibrium: European users of the designated platforms receive neither the first-party feature nor an open platform on which a rival agent could match it. This is something the Google case does not address. A reactive, request-based regime can deter a gatekeeper from shipping a deeply integrated agent in the European Union; it cannot compel that gatekeeper to ship-and-open; and it cannot prevent the outcome in which Europe is left with nothing while the same feature ships elsewhere.

Here, the logic of the model carries over. In the model's terms, the interoperability duty is triggered by offering, so it removes the gatekeeper's preferred option to offer the agent and to

keep the platform closed and forces a choice between offering-and-opening and withholding. The gatekeeper withholds when the ecosystem rent it would lose by opening exceeds the sum of its agent monetisation and the device sales it would forfeit to the rival platform by withdrawing. The decisive result is that a gatekeeper rationally withholds its agent from a regulated market only if it expects to keep its users despite offering them nothing.

Stated that way, the decision is informative about the firm that takes it. A gatekeeper withholds only if it expects to keep its users while offering them nothing, that is, only if it is confident that its ecosystem holds them for reasons unrelated to the withheld feature. As the Commission characterised it in its Article 6(7) specification decision addressed to Apple, the Case DMA.100204, end users of Apple's own services and hardware are locked into its ecosystem, a captivity the Commission illustrated through the Apple Watch and named with Apple's own term for it, the "stickiness" of the ecosystem (European Commission, 2025b). Apple, whose hardware, software and services are tightly integrated, is better placed to make that calculation than a platform whose users face a lower cost of leaving; and it is telling that Google, on Android, has instead moved toward compliance on the agent slot rather than withdrawal. But a hold that strong is precisely the market power that the interoperability regime exists to constrain. Apple's willingness to withdraw Siri AI from its European users on iOS and iPadOS is therefore not only a cost imposed by the DMA; it is evidence for the DMA's premise, since a genuinely contestable device market would have disciplined Apple into offering rather than withholding.

The question this raises is what an affected consumer can actually do. A consumer who values the integrated agent faces a trade-off: stay on the platform and forgo it, or switch to a rival platform that offers an equivalent and bear the switching cost. Unfortunately, the switching costs are pure waste, and the consumer loss is largest precisely in the high lock-in case, which is the very case in which the gatekeeper chooses to withhold.

6 Discussion

Three findings emerge from putting the model and the proceeding side by side. The first is that the hardware layer behaves as the model predicts. The Commission found preferential access to RAM in the Google proceeding, which highlights the importance of establishing a regulatory floor. The second is that the integrity carve-out, the non-discrimination standard and the case-by-case enforcement mode all fail for one reason: nothing at the hardware layer is observable from outside the gatekeeper. The remedy answers this with notification and justification alone. The third is that the obligation's reach is bounded by the gatekeeper's option not to ship at all, a margin the Regulation does not address.

Each has a policy corollary. If parity is postulated rather than measured, the natural next step is not a stricter standard but an observable one: exposing scheduling and residency decisions to independent audit would turn a rule into something a rival can check. If a stricter floor is

self-reinforcing, as Corollary 1 implies, the returns to raising it are larger than they look at the margin. And if withdrawal is the residual margin, an obligation triggered by offering will, for the most locked-in ecosystems, sometimes buy European users nothing at all.

The most immediate research path is the one this thesis had to leave open. Its case study analyses the proposed measures of April 2026, which were the operative documents at the time of writing. The final decision of 16 July 2026 postdates the analysis. More importantly, the non-confidential version of that decision, which sets out the Commission's reasoning rather than only the measures it imposes, had not been published when this research closed. That document is where the questions this thesis raises will actually be answered. It will show how the Commission understands the hardware layer it has just regulated, and in particular whether it treats the accelerator as one resource among several or as the component on which on-device inference turns. It will show what evidentiary basis supported the finding of preferential access. And, read against the April draft, it will show what the consultation changed and why. This is especilly true for the provision this thesis singled out as the weak point of the April draft developed in Section 5.2.1, the user-toggle of §108(b), under which parity in hardware allocation was to be obtained only where the user affirmatively removed the gatekeeper's preferential allocation. That provision did not survive into the final measures, which require non-discriminatory allocation without that qualification. The analysis developed in Section 5.2.1 therefore identified, before the outcome was known, the point on which the consultation turned.

The floor-sitting mechanism was real enough to shape the regulator's own first draft. A Commission writing a remedy from scratch, with the AICore finding in front of it and the whole of Article 6(7) at its disposal, initially settled on a text that preserved the gatekeeper's advantage by default and made parity an option the user had to exercise. Correcting that took a full consultation cycle and close to three months. This is the clearest available illustration of the thesis's central claim: parity at the hardware layer is not delivered by the obligation itself but has to be extracted, provision by provision, from a process that can get it wrong at the first attempt. Understanding how a single sub-paragraph came to be drafted and then withdrawn can say a lot about the practical difficulty of specifying interoperability.

7 Conclusion

This thesis asked whether the DMA reaches a gatekeeper's privileged access to the hardware on which an AI agent runs, and whether reaching it is the same thing as securing it on equal terms.

On scope, Article 6(7) extends to hardware features controlled via the operating system, and the Commission's proceeding in Case DMA.100220 applies the obligation to on-device AI in granular detail, down to hardware-resource allocation, memory residency and background execution. It documented along the way that Alphabet's AICore enjoyed preferential access to RAM until the Commission intervened.

On terms, the answer is different. The distinction between access and parity survives: an obligation that guarantees reach does not guarantee quality-equivalent reach. The remedy as first drafted left the gatekeeper's advantage in place by default, removable only by an affirmative user action. The non-discrimination standard on which it rests is hardly observable at the hardware layer. The residual gap is therefore one of enforcement mode: reactive, case-by-case specification, qualified by an integrity exception that is hardest to verify exactly where the conduct is least visible.

What the model adds is that this conduct is predictable. The model predicts that a gatekeeper facing an access obligation either supplies access at exactly the level the obligation requires and no more, or grants full parity outright where rival access adds more to the device price than its own agent earns it. For the gatekeepers at issue here, the second result is unlikely: both monetise downstream easily enough that the floor is the expected outcome. The same logic, extended, gives the revealed-preference reading of withdrawal. Apple's decision to withhold Siri AI from the European Union while shipping it where the DMA weighs lightly is simultaneously a cost of the regulation and evidence for its premise, since only a firm confident of its ecosystem lock-in can afford to offer its users nothing.

Finally, Articles 6(7) and 6(5) are complements rather than substitutes. Enforcing the downstream prohibition lowers the monetisation that drives upstream degradation, and enforcing the upstream duty shrinks the captive base that downstream steering exploits; enforcing neither lets the two compound.

The limits of the analysis are those of its material. The Google proceeding is preliminary and its measures may change; the Apple facts are recent, contested, and drawn from statements made in an ongoing dispute. The model is deliberately reduced-form and its central empirical claim that withholding reveals lock-in is an inference from structure, not a measurement of preferences. These limits indicate the research paths that follow. Empirically, once remedies bind, rival agents' realised access can be benchmarked across platforms; latency, scheduling, and memory residency turning the parity question from a postulate into a measurement. And, the final decision in DMA.100220 will show whether the specification power already being exercised can deliver parity by default and verification by measurement, the two design features whose absence this thesis has identified. The waters at the source are still thin; they are worth fishing.

References

Apple Inc. (2025). *Condensed consolidated financial statements (unaudited): Fourth quarter and fiscal year 2025*. https://investor.apple.com/

Apple Inc. (2026, June 8). *Due to DMA, Siri AI delayed in EU for iOS 27 and iPadOS 27* [Press release]. Apple Newsroom. https://www.apple.com/newsroom/2026/06/due-to-dma-siri-ai-delayed-in-eu-for-ios-27-and-ipados-27/

Bostoen, F., & Krämer, J. (2026). How future-proof is the DMA? A case study of AI agents. *Journal of Competition Law & Economics*. Advance online publication. https://doi.org/10.1093/joclec/nhag005

Bourreau, M. (2022). *DMA: Horizontal and vertical interoperability obligations* (Issue Paper). Centre on Regulation in Europe (CERRE).

Carugati, C. (2022). *How to implement the self-preferencing ban in the European Union's Digital Markets Act* (Policy Contribution 22/2022). Bruegel.

Economides, N. (1996). Network externalities, complementarities, and invitations to enter. *European Journal of Political Economy, 12*(2), 211–233.

Economides, N. (1998). The incentive for non-price discrimination by an input monopolist. *International Journal of Industrial Organization, 16*(3), 271–284.

European Commission. (2025a). *Commission Decision of 19 March 2025 specifying measures pursuant to Article 6(7) of Regulation (EU) 2022/1925* (Case DMA.100203 — Apple, connected physical devices).

European Commission. (2025b). *Commission Decision of 19 March 2025 specifying measures pursuant to Article 6(7) of Regulation (EU) 2022/1925* (Case DMA.100204 — Apple, interoperability request process).

European Commission. (2026a). *Preliminary findings and proposed measures of 27 April 2026 pursuant to Articles 6(7) and 8(2) of Regulation (EU) 2022/1925* (Case DMA.100220 — Google Android, interoperability with artificial-intelligence functionality).

European Commission. (2026b). *Statement by Commission spokesperson Thomas Regnier at the midday press briefing of June 9, 2026* [Press briefing]. Brussels, Belgium.

European Commission. (2026c). *Decision of 16 July 2026 specifying measures pursuant to Articles 6(7) and 8(2) of Regulation (EU) 2022/1925 — Final measures* (Case DMA.100220 — Google Android, interoperability with artificial-intelligence functionality). Provisional non-confidential version. https://ec.europa.eu/competition/digital_markets_act/cases/202629/DMA_100220_2683.pdf

Fletcher, A., Crémer, J., Heidhues, P., Kimmelman, G., Monti, G., Podszun, R., Schnitzer, M.,

Scott Morton, F., & de Streel, A. (2024). The effective use of economics in the EU Digital Markets Act. *Journal of Competition Law & Economics, 20*(1–2), 1–19.

Fumagalli, C., & Motta, M. (2020). Dynamic vertical foreclosure. *The Journal of Law and Economics, 63*(4), 763–812. https://doi.org/10.1086/709584

Hagiu, A., Teh, T.-H., & Wright, J. (2022). Should platforms be allowed to sell on their own marketplaces? *The RAND Journal of Economics, 53*(2), 297–327.

Hagiu, A., & Wright, J. (2025). Artificial intelligence and competition policy. *International Journal of Industrial Organization, 103*, 103134.

Heidhues, P., Köster, M., & Kőszegi, B. (2023). Steering fallible consumers. *The Economic Journal, 133*(652), 1430–1465. https://doi.org/10.1093/ej/ueac093

Hovenkamp, E. (2022). *Proposed antitrust reforms in Big Tech: What do they imply for competition and innovation?* SSRN. https://ssrn.com/abstract=4127334

Johnen, J., & Shekhar, S. (2026). Foreclosure incentives with network effects: A framework for screening digital mergers. *Journal of Competition Law & Economics*. https://doi.org/10.1093/joclec/nhag011

Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *The American Economic Review, 75*(3), 424–440.

Lee, K. J. (2021). Architecture of neural processing unit for deep neural networks. In S. Kim & G. C. Deka (Eds.), *Hardware accelerator systems for artificial intelligence and machine learning* (Advances in Computers, Vol. 122, pp. 217–245). Elsevier. https://doi.org/10.1016/bs.adcom.2020.11.001

Meuser, T., Lovén, L., Bhuyan, M., Patil, S. G., Dustdar, S., Aral, A., Bayhan, S., Becker, C., de Lara, E., Ding, A. Y., Edinger, J., Gross, J., Mohan, N., Pimentel, A. D., Rivière, E., Schulzrinne, H., Simoens, P., Solmaz, G., & Welzl, M. (2024). Revisiting edge AI: Opportunities and challenges. *IEEE Internet Computing, 28*(4), 49–59. https://doi.org/10.1109/MIC.2024.3383758

Morozovaite, V. (2023). The future of anticompetitive self-preferencing: Analysis of hypernudging by voice assistants under Article 102 TFEU. *European Competition Journal, 19*(2), 410–448. https://doi.org/10.1080/17441056.2023.2200623

Motta, M. (2023). Self-preferencing and foreclosure in digital markets: Theories of harm for abuse cases. *International Journal of Industrial Organization, 90*, 102974.

Padilla, J. (2024). Fairness and contestability in the provision of software application stores services. *Journal of Antitrust Enforcement, 12*(2), 309–314.

Padilla, J., Perkins, J., & Piccolo, S. (2022). Self-preferencing in markets with vertically integrated gatekeeper platforms. *The Journal of Industrial Economics, 70*(2), 371–395.

Peitz, M. (2022). *The prohibition of self-preferencing in the DMA* (Issue Paper). Centre on Regulation in Europe (CERRE).

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). *Official Journal of the European Union*, L 265, 1–66.

Rey, P., & Tirole, J. (2007). A primer on foreclosure. In M. Armstrong & R. Porter (Eds.), *Handbook of industrial organization* (Vol. 3, pp. 2145–2220). Elsevier. https://doi.org/10.1016/S1573-448X(06)03033-0

Salop, S. C., & Scheffman, D. T. (1983). Raising rivals' costs. *The American Economic Review, 73*(2), 267–271.

Scott Morton, F. (2024). *Entry and competition in mobile app stores* (Working Paper 03/2024). Bruegel.

Scott Morton, F., Crawford, G., Crémer, J., Dinielli, D., Fletcher, A., Heidhues, P., & Schnitzer, M. (2023). Equitable interoperability: The "supertool" of digital platform governance. *Yale Journal on Regulation, 40*, 1013–1055.

Singh, R., & Gill, S. S. (2023). Edge AI: A survey. *Internet of Things and Cyber-Physical Systems, 3*, 71–92. https://doi.org/10.1016/j.iotcps.2023.02.004

StatCounter. (2026). *Mobile operating system market share worldwide*. Retrieved July 2, 2026, from https://gs.statcounter.com/os-market-share/mobile/worldwide