Whose Data Moat? Workflow Ownership and Competitive Advantage in Vertical Artificial Intelligence, with a Research Agenda for Real Estate

Eli Bock *Independent Researcher*

Working paper — July 2026

Disclosure: the author is a practitioner in the sector studied, operating artificial-intelligence services for real-estate firms. The paper presents no data from that practice; the disclosure is made so that readers may weigh the analysis with the author's position in view.

Structured abstract

Purpose — Whether data confers sustainable competitive advantage remains contested: sceptics hold that data-based advantages are transient absent isolating mechanisms; the affirmative theory names candidate mechanisms, but only for consumer platforms. This paper asks under what conditions the workflow data generated by vertically deployed artificial intelligence constitutes a strategic resource — and to whom it accrues.

Design/methodology/approach — A conceptual synthesis drawing on three literatures — the resource-based canon, the data-as-strategic-resource debate, and real-estate technology research — with every citation verified against bibliographic databases. Integrating these with constructs from transaction-cost economics, the property-rights theory of the firm, and the information-technology outsourcing literature, the paper develops a workflow-ownership framework, operationalises the VRIN criteria as independently observable measures, and derives propositions. The framework is applied to the real-estate vertical from published sources.

Findings — Whether workflow data satisfies the resource-based criteria depends on position along four observable deployment dimensions: mode, integration depth, cross-entity linkage, and accumulation cadence. Five testable propositions state the dependencies. The most consequential is distributional: the resource accrues by default to whoever operates the workflow — typically the vendor, not the client whose activity generates it — and contract can reallocate its artefact components but not its tacit residue. Applied to real estate, the framework locates prospective advantage only in unpooled, confidential, workflow-level joins, and predicts commodity competition wherever joins have been pooled or commercialised, as the vertical's own listing institutions illustrate.

Originality/value — The paper's contribution is integrative: it brings the data-moat debate into registration with the transaction-cost, property-rights, and outsourcing literatures it has not previously met; extends data-enabled-learning theory to low-frequency professional verticals; and converts an evidenced empirical gap — the review identified no study measuring deployed-AI outcomes within real-estate firms — into a research agenda specified to the level of constructs and designs.

Keywords: artificial intelligence; resource-based view; data network effects; competitive advantage; transaction cost economics; proptech; real estate; research agenda

Paper type: Conceptual paper

1. Introduction

Whether data confers sustainable competitive advantage is a question the scholarly literature has left unfinished. One position, developed with formal precision, holds that learning from accumulating data can generate advantage with the durability of a network effect (Hagiu and Wright, 2023), building on the data-network-effects mechanism proposed for user value creation on artificial-intelligence platforms (Gregory, Henfridsson, Kaganer and Kyriakou, 2020). A sceptical position holds that data-based advantages are transient absent some further isolating mechanism, because commodity data is cheaply replicated and rarely satisfies the established criteria for a strategic resource (Lambrecht and Tucker, 2015) — a scepticism whose economic premise, the non-rivalry of data, has been formalised independently (Jones and Tonetti, 2020). The positions are logically compatible; what stands between them is an unfinished specification — which isolating mechanisms operate, in which settings, and in whose hands. The disagreement is consequential. It bears on every firm deciding whether to treat its operational data as an asset, and on every vendor whose service generates such data. And as generative artificial intelligence diffuses — its frontier capabilities available to all comers at published prices — it bears on the question of where, if anywhere, durable advantage in the artificial-intelligence era can be built at all (Azoulay, Krieger and Nagaraj, 2024).

The question is not hypothetical. Consider two provisions of the same artificial-intelligence capability for processing transaction documents. One vendor licenses it as software: the client's staff upload documents and review outputs. Another operates the workflow end-to-end across a portfolio of clients. The underlying model is identical; the accumulating positions are not — and, this paper will argue, only one of them can satisfy the established criteria for a strategic resource. Which one, and why, is what the data-moat debate has left unspecified.

The debate has been conducted, however, on a narrow evidentiary base: formal models on the one hand, and illustrations drawn from high-frequency consumer platforms on the other. Whether its mechanisms operate in the professional verticals where most economic activity occurs — low-frequency, high-stakes, relationship-mediated settings — is neither theorised nor tested. Real estate is a revealing case in point. Its workflows are episodic, document-intensive, and fragmented across organisational boundaries (Saull, Baum and Braesemann, 2020; Braesemann and Baum, 2020) — on every dimension the opposite of the platform settings from which the affirmative theory was derived — and and its research literature, though attentive to technology adoption, contains no study identified by the present review that measures the outcomes of artificial-intelligence systems deployed within its firms, and none treating the data such deployments generate as a competitive resource (Section 3.3). A theory dispute that turns on deployment-level contingencies thus confronts a vertical whose deployment-level evidence, so far as could be established, does not exist.

This paper addresses both deficits — the theoretical impasse and the empirical silence — through three moves. The first is a synthesis: an integrative review, conducted against bibliographic databases under a documented verification protocol, assembles the resource-based canon, the data-as-strategic-resource debate, and real-estate technology research, and locates four specific points of tension among them. The second is analytical: the paper argues that the data-moat controversy persists because it has been conducted over an under-specified unit of analysis, and develops a framework that shifts assessment from data as a stock to the *data-generating workflow* — an organisationally embedded process whose ownership, integration depth, cross-entity linkage, and accumulation cadence jointly determine whether its deposit satisfies the resource-based criteria of value, rarity, imperfect imitability, and non-substitutability (Barney, 1991), read through the logic of complementary assets (Teece, 1986). From this framework five testable propositions are derived, including a distributional claim the platform-derived literature cannot raise: that in vendor-operated deployments the data resource accrues by default to the operator rather than to the client whose activity generates it. The third move converts the vertical's evidentiary silence into a programme: the framework's constructs are mapped onto real estate's documented workflow characteristics, and a research agenda assigns each proposition to the empirical design capable of testing it.

The paper's contributions are accordingly threefold. To strategy research, it offers a reconciliation of the data-moat controversy in which the sceptical and affirmative positions become poles of a single measurable construct rather than contradictory claims about the nature of data. To the theory of data-enabled advantage, it supplies an extension beyond platform settings, with mechanisms definable at the scale of a single professional relationship (whether they bind at that scale depends on the workflow's event cadence, cross-entity linkage substituting for scale in low-cadence verticals), and which therefore reaches the verticals the existing theory does not. To real-estate research, it provides what the present review indicates is the first resource-based treatment of workflow data in the vertical, together with an agenda whose designs are specified to the level of constructs and are, in the case of its first design, executable by a single investigator with field access.

The remainder of the paper proceeds as follows. Section 2 documents the review method. Section 3 synthesises the three literatures and derives the four tensions that motivate the analysis. Section 4 develops the workflow-ownership framework and operationalises the resource-based criteria for workflow-generated data. Section 5 states the propositions. Section 6 applies the framework to the real-estate vertical, and Section 7 sets out the research agenda. Section 8 discusses contributions, implications for practice, and limitations.

2. Review method

This is a conceptual paper, and its sources were selected as theory syntheses select them: to represent the distinct positions in the literatures it joins — the resource-based tradition in strategic management, the contemporary debate over data as a source of competitive advantage, the governance literatures on specificity, incomplete contracts, and outsourcing, and empirical research on technology and artificial intelligence in real estate. One discipline was nonetheless imposed throughout. Candidate works were identified through structured keyword search of the OpenAlex and Crossref bibliographic databases (June–July 2026), supplemented by citation chaining from anchor works, and every source was verified before inclusion: a work entered the corpus only where its digital object identifier — or, for monographs, its publisher catalogue record — resolved to canonical metadata whose authorship and publication year matched the citing claim. The corpus comprises forty-two works: six foundational contributions to the resource-based view and dynamic-capabilities tradition, six methodological works, twelve contributions to the data-as-strategic-resource debate, five from the transaction-cost, property-rights, and outsourcing literatures, twelve studies of technology and artificial intelligence in real estate, and one study of the vertical's listing institutions.

Selection favoured well-cited works, deliberately included the strongest available statements of positions the paper argues against, and weighted recency in the contemporary bodies. One boundary warrants disclosure. A portion of the corpus was engaged through verified metadata and published abstracts rather than full text, where publisher access restrictions applied; those works are identified in the corpus documentation accompanying the paper, and no fine-grained textual claims rest on them. The absence claims of Section 3.3 are bounded by this method.

The synthesis that follows is organised around four points of tension identified during the review — two internal to the data-moat debate, and two arising between literatures. The paper's analytical contribution, developed in Section 4, proceeds directly from the diagnosis of these tensions.

3. Three literatures in tension

3.1 The resource-based canon

The resource-based view explains persistent performance differences between firms by the resources they control rather than the product-market positions they occupy. Wernerfelt (1984) proposed the inversion of perspective: firms may be analysed as bundles of resources, with resource position barriers playing the role that entry barriers play in market-based analysis. Barney (1991) supplied the framework through which the tradition is now most commonly applied, specifying four conditions a resource must satisfy to sustain competitive advantage: it must be valuable, rare among competitors, imperfectly imitable, and non-substitutable. Peteraf (1993) consolidated the tradition into four cornerstones — heterogeneity, ex post limits to competition, imperfect mobility, and ex ante limits to competition — of which imperfect mobility carries particular weight in what follows: a resource that cannot be traded, or that loses value when separated from the firm in which it is embedded, cannot be bid away.

The criteria are not independent, and their interrelations matter for what follows. Value without rarity describes competitive parity: a resource all rivals hold may be necessary to compete but differentiates no one. Rarity without imperfect imitability describes a head start: a resource rivals lack today but can acquire tomorrow yields advantage only for the duration of the acquisition lag. It is the third criterion that carries the durability burden, and Barney (1991) locates its satisfaction in identifiable isolating conditions — unique historical circumstances, causal ambiguity, and social complexity — each of which describes a way in which the path by which a resource was accumulated resists compression. Peteraf's (1993) treatment adds the market-side counterpart: even a valuable, rare, hard-to-imitate resource sustains no advantage if it can be bid away or if its acquisition cost capitalises the rents it will earn; hence the weight placed below on imperfect mobility, the condition that a resource loses value when separated from the context in which it is embedded. These interrelations discipline any claim of the form "X is a strategic resource": the claim must specify not merely that X is useful but why its accumulation path resists compression and why it cannot be traded away from its context — requirements the data-moat debate, reviewed next, has largely engaged one criterion at a time.

Two extensions of the canon bear directly on the question at hand. Teece (1986) addressed why innovators frequently fail to profit from their own innovations while imitators and asset-holders profit instead: where the appropriability regime around an innovation is weak, value migrates to the holders of specialised complementary assets — the capabilities, channels, and co-specialised infrastructure required to convert the innovation into delivered value. The argument was formulated for an earlier generation of technology, but its structure is general, and it supplies the interpretive frame adopted in Section 4: where the technological innovation itself is widely accessible, the locus of appropriation must be sought in whatever complements it. Second, the dynamic-capabilities extension (Teece, Pisano and Shuen, 1997; Teece, 2007) addresses how advantage is sustained under technological change: not by holding a stock of resources, but through the organisational capacity to integrate, build, and reconfigure them — disaggregated by Teece (2007) into the microfoundations of sensing, seizing, and transforming. On this view, a data position is not itself a durable advantage; the capability that continually renews and redeploys the position is.

3.2 The data-moat debate

Whether data satisfies the resource-based criteria is now contested across strategy, information systems, and economics. The sceptical position comes first, because it is the position any affirmative claim must survive.

Lambrecht and Tucker (2015) argue that big data fails the resource-based test on nearly every criterion: data is non-rivalrous, cheaply replicated, widely available from multiple sources, and rarely inimitable, and any advantage it confers is therefore transient absent some further isolating mechanism. Jones and Tonetti (2020) formalise the economic property on which much of this scepticism rests: data's non-rivalry implies that the same information can, in principle, be used simultaneously by any number of firms, so that raw data constitutes a weak barrier and the economically consequential questions concern ownership and exclusivity arrangements rather than the data itself. Goldfarb and Tucker (2019), surveying digital economics, supply the framing observation that digitisation lowers the costs of search, replication, and verification — with the implication that where information itself becomes cheap, advantage migrates to whatever complements it. Within management theory, Clough and Wu (2020) temper the strongest data-based claims from a different direction: even where data-driven learning operates, its advantages are bounded by the structure of the surrounding ecosystem, by data spillovers, and by the decentralisation of complementors over whom the focal firm exercises no control.

The sceptical position is strengthened, rather than weakened, by noting what it does not claim. Lambrecht and Tucker (2015) do not deny that data is useful; their argument is the disciplined resource-based one that usefulness is the weakest of the criteria, and that the properties which would carry durability — rarity and imperfect imitability — are exactly the properties commodity data lacks. Nor does the non-rivalry result of Jones and Tonetti (2020) depend on data being freely shared in practice; it establishes that nothing in the *nature* of data prevents its simultaneous use by all comers, so that whatever prevents such use must be sought in institutions — property arrangements, contracts, positions — rather than in the asset itself. Stated this way, the sceptical position issues a precise challenge that any affirmative account must answer: name the institutional mechanism that makes a particular body of data excludable and its accumulation path incompressible. It is a challenge the present framework takes up directly in Section 4.

The affirmative position holds that these properties, while real, do not exhaust the question. Gregory, Henfridsson, Kaganer and Kyriakou (2020) introduce the construct of data network effects: where artificial-intelligence capability is applied to data that accumulates through platform use, the value of the platform to each user can grow with the size of the user base through a mechanism distinct from conventional network externalities. Hagiu and Wright (2023) supply the most precise statement of the conditions under which such mechanisms yield durable advantage, distinguishing within-customer from across-customer learning and deriving formally when data-enabled learning generates network-effect-like dynamics rather than transient head starts. Farboodi and Veldkamp (2021) model the macro-structure of the same feedback: data arises as a by-product of economic activity, feeds back into the productivity of the firm that captures it, and thereby compounds. Most directly pertinent to the present question, Azoulay, Krieger and Nagaraj (2024) argue that competitive advantage in generative artificial intelligence will rest on classical moats — complements, distribution, and proprietary data embedded in workflows — rather than on access to frontier models, whose capabilities diffuse rapidly across providers.

A third stream, developed within information systems, mediates between the two positions. Gupta and George (2016) develop the construct of big data analytics capability from resource-based foundations, finding that tangible data resources produce value only in combination with human and intangible organisational resources. Fosso Wamba, Gunasekaran, Akter, Ren, Dubey and Childe (2016) provide large-sample evidence that the performance effects of such capability operate primarily through process-oriented dynamic capabilities, and Mikalef, Pappas, Krogstie and Giannakos (2017), reviewing the literature systematically, conclude that the relationship between data assets and competitive advantage documented in empirical work is mediated rather than direct. Krakowski, Luger and Raisch (2022) add an instructive boundary result from an unusually clean empirical setting: where artificial intelligence substitutes for capabilities that previously differentiated firms, those capabilities cease to be sources of advantage, and advantage re-forms around the complementarities between human and machine — a resource-based analysis in which the technology itself is competitively neutral and the configuration around it is not.

The mediated-capability stream needs more than its usual one-line summary, because its empirical weight is considerable and its implication is routinely understated. Gupta and George (2016) do not find that data fails to matter; they find that the construct which predicts performance is a *bundle* — tangible data and technology resources, human analytical skill, and intangible elements including a data-driven culture and organisational learning — and that the bundle's components are individually insufficient. Fosso Wamba et al. (2016), across a sample large enough to discipline the claim, locate the operative path from analytics capability to performance in process-oriented dynamic capabilities: the effect of the asset runs through the routines that deploy it. Mikalef et al. (2017), surveying the accumulated evidence systematically, conclude that direct data-to-advantage effects are scarcely documented anywhere in the literature, mediated effects nearly everywhere.

The relocation this stream has already performed is the point. Its constructs have, in effect, moved the unit of analysis: what its constructs describe are properties of organisational processes, not of datasets. What it has not done — because its instruments are firm-level surveys — is ask *who owns* the process when generation and exploitation are split across a vendor boundary. That question is invisible at the survey level. It is the one the present framework places at the centre.

Two tensions internal to this debate structure everything that follows, and it is worth stating each with precision, because neither is a flat contradiction. The first — denote it **C1** — is a specification problem left open between the positions. The sceptical argument concludes that data-based advantages are transient *absent further isolating mechanisms*; the affirmative theory supplies candidate mechanisms — learning structures and accumulation feedback (Gregory et al., 2020; Hagiu and Wright, 2023). The positions are therefore logically compatible, and the live question between them is not whether isolating mechanisms can exist but *which* mechanisms operate, in which settings, and in whose hands — a specification the formal models state in terms of learning structure while remaining silent on the organisational and institutional arrangements that instantiate it outside consumer platforms. The second — **C2** — divides those who locate the operative resource in the accumulating data itself from those who locate it in the organisational capability that exploits it (Gupta and George, 2016; Fosso Wamba et al., 2016). It is argued below that both tensions remain open in part because the debate has been conducted over an under-specified unit of analysis — and that completing the specification requires constructs the strategy-of-data literature has so far not imported.

3.3 Artificial intelligence and real estate: the evidentiary stage

Against the theoretical density of the preceding debate, the literature on artificial intelligence in real estate is empirically thin in a specific and consequential way. Its most visible contributions are reviews and frameworks. Ullah, Sepasgozar and Wang (2018) provide the most-cited systematic review of smart real-estate technology, cataloguing adoption drivers and barriers across a literature composed predominantly of further reviews and surveys. Starr, Saginor and Worzala (2020) map the waves of artificial intelligence, the internet of things, and big data onto real-estate practice in conceptual terms, and Souza, Koroleva, Worzala, Martin, Becker and Derrick (2020) consider the implications of automation for real-estate investment organisations at the level of organisational strategy. A second stratum measures attitudes and readiness rather than deployment: valuers' receptiveness to artificial intelligence in valuation (Abidoye and Chan, 2017) and, most recently, professional readiness and organisational barriers to proptech adoption (Banaitienė and Cataldo, 2025) — work whose recency underscores that the field's empirical frontier remains at the stage of intention rather than outcome.

Firm-level empirical work exists but is scarce and uniformly qualitative. Vigren, Kadefors and Eriksson (2022), interviewing across the Swedish real-estate ecosystem, document why incumbent owners struggle to internalise digital technology, locating the difficulty in absorptive capacity and innovation capabilities. Saull, Baum and Braesemann (2020) examine the transaction workflow itself, asking through interview evidence whether digital technologies can accelerate it — among the few studies to take the workflow as the object of analysis, though it precedes the current generation of artificial-intelligence systems. Braesemann and Baum (2020) analyse the proptech firm landscape to ask whether real estate is becoming a data-driven market, bridging toward the data-as-resource debate but at industry rather than firm level. Battisti, Shams, Sakka and Miglietta (2019) connect big-data use to corporate real-estate processes; the vendor side and the brokerage level have also begun to receive qualitative attention (Tagliaro, Pomè, Migliore and Danivska, 2024; Asensio-Soto and Navarro-Astor, 2022; 2025).

The supply side of the vertical's technology market has begun to receive attention of its own. Tagliaro, Pomè, Migliore and Danivska (2024) study how proptech firms shape innovation in the sector, finding vendors acting as intermediaries who translate general technology into sector-specific application — evidence, from the vendor side, of exactly the conversion work that the complementary-assets logic predicts should be where value concentrates. The brokerage studies of Asensio-Soto and Navarro-Astor (2022; 2025) carry a complementary finding: digitalisation has reorganised rather than eliminated the human intermediary, whose persistence the authors attribute to the relational and exception-handling content of the work. Both findings are congenial to the framework developed below — the first because it locates the vertical's innovation activity in vendor-operated conversion, the second because relational, exception-rich work is precisely the standardisation-resistant regime in which co-adaptation carries competitive weight — but neither study poses the resource question: what accumulates, and to whom, when the intermediary's workflow becomes the site of machine execution.

What the present review could not locate in this literature is notable precisely because of what the preceding debate requires. No study identified by the review measures the outcomes of artificial-intelligence systems actually deployed within real-estate firms (the bounds of this claim are those of Section 2). None examines the deployment of generative or agentic systems in real-estate practice. And none treats the data generated by such deployments as a competitive resource — despite the fact that the vertical's defining characteristics (episodic, high-stakes transactions; document-intensive processes; data fragmented across brokerages, lenders, property managers, and public records) bear directly on the conditions the affirmative theory specifies. Two further tensions follow. **C3**: the affirmative theory of data-enabled advantage was developed for high-frequency consumer platforms with millions of users, and whether its mechanisms operate in low-frequency, relationship-mediated professional verticals is neither theorised nor tested. **C4**: the strongest contemporary statement of the thesis examined here — that generative-AI advantage rests on workflows and proprietary data rather than model access (Azoulay et al., 2024) — is argued conceptually, and the vertical literature that could evidence it produces none of the required observation.

3.4 Diagnosis: an under-specified unit of analysis

The four tensions are collected in Table 1; the diagnosis they share follows.

Table 1. Four tensions in the reviewed literatures.

Tension	Between	One-line statement	Resolved

C1	Sceptical and affirmative data-moat positions	An unfinished specification: which isolating mechanisms, in which settings, in whose hands	§4.4
C2	Data-as-resource and capability streams	Whether the operative resource is the accumulating data or the capability exploiting it	§4.4
C3	Platform theory and professional verticals	Whether data-enabled-learning mechanisms operate outside high-frequency consumer platforms	§4.4, §6
C4	Industry-level conjecture and deployment evidence	The strongest contemporary thesis is argued conceptually; the vertical literature supplies no deployment observation	§5, §7

The four tensions admit a common diagnosis. The properties invoked across C1 do not attach to the same object: non-rivalry and replicability are properties of data as a *stock* — a dataset that, once constituted, can in principle be copied and used elsewhere — while accumulation and learning feedback are properties of a *generating process* — an ongoing activity that produces observations as a by-product and improves as it consumes them (Farboodi and Veldkamp, 2021). The capability stream underlying C2 has, as noted above, already made this relocation implicitly: its constructs describe configured organisational processes, not datasets. The diagnosis, then, is not that the literature lacks the stock–process distinction, but that it has left the distinction's consequences unspecified in two respects. First, no account in the debate individuates the *workflow* — the recurring, organisationally embedded process — as the unit against which the resource-based criteria are assessed, which is why the sceptics' open clause ("absent isolating mechanisms") and the theorists' candidate mechanisms have not been brought into registration. Second, and more consequentially, no account in this debate asks who *owns* the generating process when generation and exploitation are split across a vendor boundary — the configuration through which artificial-intelligence capability actually enters most professional verticals. That question has been studied, for an earlier technological generation, by the literatures on transaction-cost economics, incomplete contracts, and information-technology outsourcing. Section 4 imports their constructs and integrates them with the data-enabled-learning mechanisms. The re-posed question — under what conditions does ownership of a data-generating workflow satisfy the resource-based criteria — then has a determinate answer.

4. The framework: workflow ownership as the boundary condition

4.1 Shifting the unit of analysis

Section 3 closed on a two-part diagnosis: the stock–process distinction is implicit in the existing debate but has not been carried through to the workflow as a unit of analysis, and the ownership question that arises when generation and exploitation split across a vendor boundary has not been asked within the strategy-of-data literature at all. The present section develops the framework that completes this specification. Its contribution is deliberately integrative: the constructs it assembles are drawn from transaction-cost economics (Williamson, 1979), the property-rights theory of the firm (Grossman and Hart, 1986; Hart and Moore, 1990), the information-technology outsourcing literature (Levina and Ross, 2003; Lacity, Khan and Willcocks, 2009), and the data-enabled-learning models (Hagiu and Wright, 2023) — literatures that have not previously been brought to bear jointly on the question of data as a strategic resource. The claim is not that any single element is new; it is that their integration, operationalised for machine-executed workflow data, yields determinate and testable answers where the data-moat debate has produced open clauses.

The framework's object is the *data-generating workflow*: the recurring, organisationally embedded process whose execution produces data as a by-product (Farboodi and Veldkamp, 2021). Taking the workflow as the unit of assessment preserves what is correct in both of Section 3's positions. The sceptics are correct about stocks: a dataset, once constituted, is non-rivalrous, and where comparable data can be purchased, scraped, or reconstructed, holding it confers at most a transient head start. The affirmative theorists are correct about processes: data-enabled learning and data network effects are not claims about holding data but about the coupling of an ongoing activity to an improving system, operative only while generation continues. The open clause between them — *which isolating mechanisms, operating where, in whose hands* — is answered below by specifying the organisational location of the generating activity: who executes the workflow, in whose systems its by-product accrues, and on what terms. One question remains once that location is specified. *Under what conditions does ownership of a data-generating workflow satisfy the resource-based criteria?*

The shift also aligns the debate with the older logic of complementary assets. In Teece's (1986) account, an innovation whose appropriability regime is weak yields its value to the holders of specialised complementary assets required for commercialisation. Contemporary artificial intelligence presents an unusually pure instance of a weak appropriability regime: frontier model capabilities diffuse rapidly across competing providers and are accessible to all comers at published prices (Azoulay, Krieger and Nagaraj, 2024). On the complementary-assets logic, nothing durable can be built on model access itself; the candidate loci of appropriation are the assets that convert general model capability into delivered value within a specific setting. The data-generating workflow is precisely such an asset, and its degree of specialisation — how deeply it is co-adapted to the client organisation it serves — is, as developed below, the principal determinant of whether the resource-based criteria are met.

4.2 The workflow-ownership construct

Workflow ownership is defined here as the degree to which a focal actor controls the execution of a data-generating workflow, as distinct from supplying tools with which another actor executes it. The construct is continuous rather than binary, but its poles are usefully distinguished. In *tool-mode* provision, the vendor supplies capability — software, a model interface, an application — and the client executes the workflow.

Tool-mode vendors are not blind, and the framework does not pretend they are. Contemporary software vendors instrument their products extensively, so three holdings must be distinguished from ownership as defined here. The first is *observation* of usage — telemetry, logs, interaction records — which tool vendors commonly hold at scale. The second is *exploitation rights* over what is observed, which are constrained by contract, privacy regulation, and client sensitivity, and are commonly confined to aggregate product improvement; that characterisation of prevailing terms is offered as description, and its systematic documentation is assigned to the research agenda. The third is *process structure*, which in tool-mode remains the client's: the vendor observes usage of its instrument, not the client's workflow, whose sequence, exceptions, and surrounding context lie outside the instrument's view. Telemetry is real accumulation. What it lacks, relative to the operated pole, is structure, completeness, and exploitation freedom.

In *operated-mode* provision, the vendor executes the workflow on the client's behalf as a service. Execution passes through the vendor's systems; the by-product accrues there, structured by the vendor's own process design; and its exploitation in service delivery is the engagement's premise rather than a permission to be negotiated. Between the poles lie configurations in which execution, observation, and exploitation rights are divided — the modal case in practice, to which the propositions return. Figure 1 collects the construct: the two poles, the four dimensions along which position is measured, and the region of the space in which the resource-based criteria — operationalised in Section 4.3 — become jointly attainable.

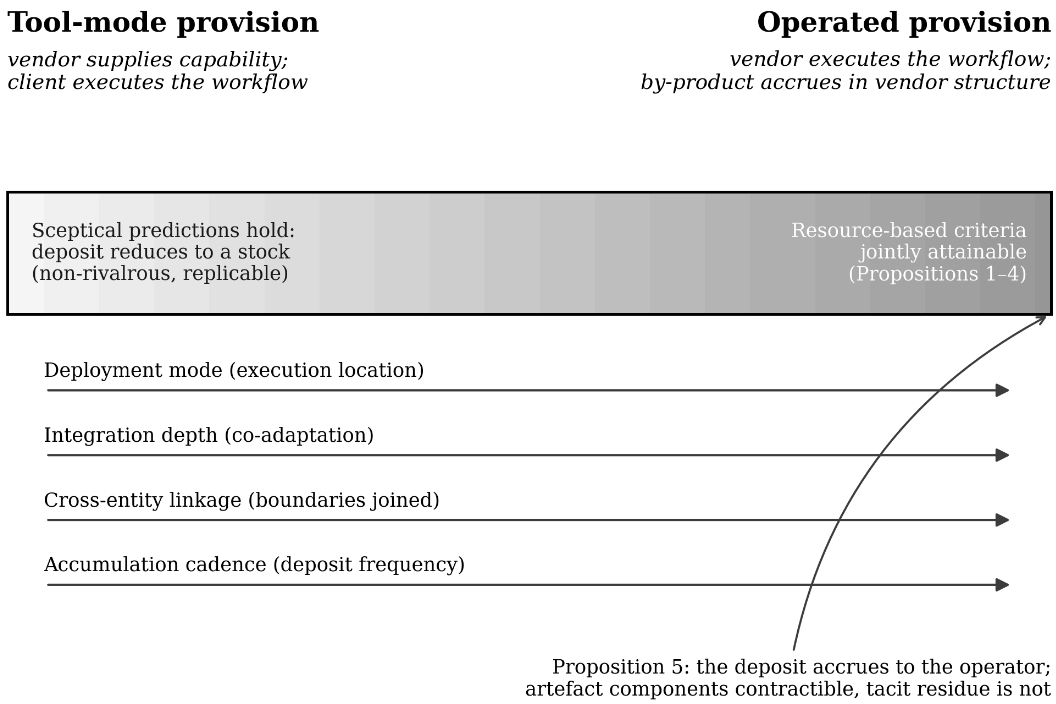

Figure 1. The workflow-ownership construct: poles, dimensions, and the region of joint criterion satisfaction.

Four dimensions give the construct empirical content. Each is grounded in an existing theoretical mechanism, and each is observable in principle at deployment level — a property exploited by the research agenda of Section 7.

Deployment mode is the location of execution itself, as defined above. It determines to whom the by-product accrues by default and is the dimension on which the vendor–client allocation of the prospective resource principally turns.

Integration depth is the extent to which the workflow is co-adapted to the particular client organisation — its entities, conventions, exception cases, and adjacent processes. The dimension is a direct application of asset specificity (Williamson, 1979) and of the specialisation of complementary assets (Teece, 1986): a deeply integrated workflow is co-specialised, valuable in conjunction with the particular client relationship and costly to redeploy, which is what renders the resulting position imperfectly mobile in Peteraf's (1993) terms rather than tradable. The outsourcing literature documents the phenomenon for human-executed services: vendors develop client-specific and cross-client capabilities that constitute their value proposition and that clients cannot readily internalise (Levina and Ross, 2003), and the accumulated evidence of that literature records persistent client-side challenges in appropriating or replacing vendor capability at termination (Lacity, Khan and Willcocks, 2009). What distinguishes the machine-executed case, developed under Proposition 5, is that a portion of the accumulated asset takes the form of transferable artefacts rather than personnel-embodied expertise — a difference that changes, without eliminating, the contractibility of the position.

Cross-entity linkage is the number and type of distinct organisational or transactional boundaries the workflow's execution spans — accounts, properties, counterparties, legal entities — measured from the deployment's observable scope, without reference to whether the resulting joins are obtainable elsewhere. The definition is deliberately neutral on rarity: whether a high-linkage workflow's deposit is in fact unobtainable from commercial or public sources is the separate, independently measurable question that Proposition 3 makes an empirical claim about, and the two can come apart — as they have, historically, wherever a vertical's high-value joins were subsequently pooled or commercialised. Linkage is the dimension least reducible to a stock: a rival in possession of each constituent dataset would still lack the joins produced in execution. It is also the dimension on which across-customer learning, the condition Hagiu and Wright (2023) identify as most conducive to durable data-enabled advantage, becomes available in professional settings whose per-client event frequency is otherwise low.

Accumulation cadence is the frequency and regularity with which workflow execution deposits new observations. Cadence governs the pace of the feedback loop that the data-economy models place at the centre of data-driven growth (Farboodi and Veldkamp, 2021), and it conditions how quickly a position, once established, compounds beyond the reach of later entrants.

The dimensions are analytically distinct; whether they are empirically correlated is itself a question for the field, and it is stated here as a conjecture rather than a premise. The conjectured direction is self-reinforcing: operated-mode provision would tend to deepen integration (the operator, executing daily, accumulates precisely the exception-knowledge that co-adaptation consists of), deep integration to widen linkage (an operator trusted with one boundary is the natural candidate for the adjacent one), and widened linkage to raise effective cadence (events sparse per entity are dense across them). If the conjecture holds, the construct describes not four independent settings but a gradient along which established operated positions migrate toward the configurations in which the resource-based criteria bind most strongly; if it fails — if the dimensions vary independently in the field — the framework loses none of its assessability but sheds the compounding dynamic. Nothing in the conjectured gradient is automatic in either case: each reinforcing step requires the client's continued delegation, which is itself a governance decision — the hinge on which the allocation question of Proposition 5 turns — and each requires of the operator the dynamic capabilities the extended resource-based view describes (Teece, Pisano and Shuen, 1997; Teece, 2007): sensing the adjacent boundary, seizing the delegation, and reconfiguring the workflow to span it. The longitudinal design of Section 7 is specified to observe the dimensions separately for exactly this reason.

A concrete illustration fixes the dimensions without resort to proprietary evidence — the two provisions of Section 1, now in full. Consider two stylised provisions of the same underlying capability — an artificial-intelligence system for processing transaction documents. In the first, a software vendor licenses the system; the client's staff upload documents and review outputs; the vendor holds usage telemetry whose exploitation is constrained by contract and whose view ends at the instrument's edge, while the workflow's surrounding steps deposit nothing structured anywhere. In the second, a services vendor operates document processing end-to-end across a portfolio of clients: intake, extraction, exception resolution, and hand-off run through the vendor's process, which is adapted to each client's document conventions and which joins, in the vendor's systems, records that no single client's perimeter contains. The underlying model may be identical in both provisions. Every framework dimension differs: mode (tool against operated), integration (generic interface against per-client co-adaptation), linkage (none against cross-portfolio joins), cadence (episodic client usage against the vendor's continuous multi-client flow). The framework's claim, developed through the criteria below, is that these differences — not any property of the model or of "data" in the abstract — determine whether a strategic resource forms, and in whose hands.

4.3 The resource-based criteria, operationalised for workflow data

The framework's second component operationalises each of Barney's (1991) criteria for the specific case of workflow-generated data. The operationalisations are stated so as to be assessable, in principle, against deployment-level observation; they are summarised in Table 2.

Value. Workflow data is valuable where its exploitation measurably improves the execution of the workflow that generates it, or of adjacent workflows — reduced cycle time, reduced error or exception rates, improved conversion or utilisation — rather than merely being voluminous. This operationalisation follows the capability stream's finding that data confers value only through process-level exploitation (Gupta and George, 2016; Fosso Wamba et al., 2016): the criterion is performance in the generating process, not possession.

Rarity. Workflow data is rare where the joins and observations it contains are not obtainable from commercial or public sources. The criterion is measured independently of the linkage dimension, by documented survey of the vertical's data market: whether any broker, platform, or pooled arrangement sells or shares a functional substitute for the deposit in question. Rarity so assessed is a property of market position rather than of scale — a modest corpus may be rare while a vast corpus assembled from purchasable feeds is not — and it is defeasible: a join that is rare at one date may be pooled or commercialised at a later one, an outcome the substitute survey detects and Proposition 3 treats as disconfirming evidence rather than as an excluded case.

Imperfect imitability. Workflow data is imperfectly imitable where a competent rival — including the client itself, acting without the incumbent operator — cannot reach performance parity in the workflow within the period over which the asset retains its value. The operationalisation is deliberately an *outcome* measure, observable independently of the framework's dimensions: the imitation lag is the elapsed time (and cost) for a displacing rival or an internalising client to match the incumbent's process performance, and it can be measured wherever displacement or internalisation is attempted — without reference to how deep the incumbent's integration was. The measure incorporates time because workflow data is the deposit of an ongoing process: its reconstruction requires not copying but re-execution. Whether the imitation lag in fact lengthens with integration depth is an empirical claim, stated as Proposition 2, not a definitional consequence; the framework would be disconfirmed on this point by observed displacements in which deeply integrated positions were matched as quickly as shallow ones.

Non-substitutability. Workflow data is non-substitutable where general-purpose capability — in the present context, access to frontier models, however capable — cannot deliver the same process-level performance without it. The criterion operationalises the distinction on which the entire framework rests: where a general model prompted without organisational context matches the performance of the workflow-data-equipped system, the data asset is substitutable and no moat exists, whatever its rarity.

Table 2. The VRIN criteria operationalised for workflow-generated data.

Criterion	Operationalisation	Principal determining dimension
Value	Exploitation measurably improves execution of the generating or adjacent workflows	Accumulation cadence; integration depth
Rarity	Contents and joins unobtainable from commercial or public sources	Cross-entity linkage
Imperfect imitability	Rival's or client's measured lag (time, cost) to performance parity upon displacement or internalisation	Integration depth; accumulation cadence
Non-substitutability	General model capability without the asset cannot match process-level performance	Integration depth; cross-entity linkage

Note: the dimension–criterion linkages in the third column are the propositions' empirical content (Section 5), not definitional consequences of the operationalisations.

4.4 What the framework resolves — and how it differs from adjacent accounts

Restated through the framework, the two internal tensions of Section 3 cease to be contradictions. C1 — whether data can constitute a moat — becomes a question of position on the workflow-ownership dimensions: the sceptical predictions hold at the tool-mode, shallow-integration, low-linkage pole, where the asset reduces to a stock with the economic properties Lambrecht and Tucker (2015) describe; the affirmative predictions become available toward the operated-mode, deep-integration pole, where the asset is inseparable from a generating process that rivals must re-execute rather than copy. C2 — data versus capability — dissolves because the workflow is jointly constituted of both: the construct's dimensions are organisational properties of a process, not attributes of a dataset, so the framework sides with the capability stream's mechanism while retaining the data asset as the object appropriated. The allocation question that the construct's deployment-mode dimension raises — the by-product of an operated workflow accrues by default to the *operator*, not to the client whose business generates it — is, to the present argument, the most consequential and least examined implication of the analysis, and it is taken up as a contractual and governance matter in Sections 6 and 8.

The framework's position relative to the accounts it integrates warrants precise statement. Relative to Hagiu and Wright (2023), whose model specifies learning structures *within* a firm that already holds customer data, the framework addresses the prior question of *which party comes to hold the data at all* in vendor-mediated verticals, and locates the answer in the organisation of the workflow rather than the structure of learning. Relative to Gregory et al. (2020) and the platform setting generally, it does not require platform scale or user network structure: its mechanisms are definable at the scale of a single professional services relationship — whether they bind at that scale depends on cadence, with linkage substituting for scale where per-entity events are sparse — which is what renders the framework applicable to the low-frequency verticals (the C3 tension) that platform-derived theory does not reach. Relative to the outsourcing literature, which established for human-executed services that vendors accumulate inappropriable cross-client expertise (Levina and Ross, 2003) and that contractual provisions imperfectly protect clients at termination (Lacity, Khan and Willcocks, 2009), the framework asks what changes when the accumulating asset is generated by machine execution: the by-product is now partly an *artefact* — structured data, learned configurations — whose transfer can be specified in ways personnel-embodied expertise never could, so that the incomplete-contracts logic (Grossman and Hart, 1986; Hart and Moore, 1990) applies with a different boundary between the contractible and the residual. The distributional analysis of Proposition 5 is, in these terms, an application of residual-control reasoning to a new asset class rather than a discovery of the underlying logic. Finally, relative to Azoulay et al. (2024), who argue at the industry level that generative-AI advantage will rest on classical moats including workflow-embedded data, the framework supplies the firm-level analytical machinery the argument presupposes: constructs, operationalised criteria, and — in the following section — testable propositions.

5. Propositions

The framework yields five propositions. Each is stated so as to be empirically testable — its constructs are the observable dimensions of Section 4.2 and its predicted outcomes are assessable against the operationalised criteria of Section 4.3 — and each is followed by the strongest objection the reviewed literature supplies, together with the answer the framework gives or the scope condition it concedes.

Proposition 1 (Mode dominance). *The deposit of operated provision satisfies the resource-based criteria jointly more often than the deposit of tool-mode provision of the same underlying capability.*

The claim is comparative and monotonic, not one of strict necessity, and it extends across the construct's dimensions: joint criterion satisfaction should rise with position, so intermediate, shared-execution configurations — the modal case in practice — should show partial satisfaction, concentrated in whichever criteria their held dimensions govern (Table 2). The reasoning: tool-mode telemetry is genuine accumulation but is structured by instrument usage rather than process design, blind to the workflow's off-instrument steps, and constrained in exploitation by contract and privacy obligation (Section 4.2); operated provision joins generation, structure, and exploitation freedom in one actor, closing the feedback loop of the data-economy models (Farboodi and Veldkamp, 2021). Two objections deserve statement. The sceptics' replication argument — whatever the operator accumulates, a rival can accumulate in parallel (Lambrecht and Tucker, 2015) — concedes the mechanism: parallel accumulation requires the rival to hold an operated position of its own, which is the claim that position, not data, is the resource. The sharper objection is the telemetry counterexample: a tool vendor whose instrumentation satisfies all four criteria jointly would refute the proposition's comparative claim, and the proposition is stated so that such a case — identified through Design 1's sampling of tool-mode deployments — would disconfirm it rather than be defined away.

Proposition 2 (Time compression). *The imitation lag — the time and cost a displacing rival or an internalising client requires to reach performance parity — increases with the incumbent position's integration depth.*

The variables are measured independently: depth ex ante, from the deployment's observable co-adaptation (client-specific process artefacts, exception coverage, adjacent-process couplings); lag ex post, from displacement and internalisation events. The predicted mechanism is time-compression: a displacing rival must reach comparable co-adaptation before its own accumulation begins, during which the incumbent's asset compounds — asset specificity (Williamson, 1979) expressed dynamically. The proposition is disconfirmable because its variables are separately observable: deployments coded as deeply integrated ex ante whose positions are matched as quickly as shallow ones upon displacement refute it, and Design 3 is specified to capture precisely such events. The strongest objection extends the ecosystem-bounds logic of Clough and Wu (2020): where architectures modularise and interfaces standardise — the conditions under which, on their account, data-driven learning advantages dissipate through the ecosystem — the co-adaptation period should collapse. The framework concedes this as a genuine scope condition — the proposition holds where workflows resist standardisation (exception-rich, relationship-mediated, document-heavy processes) and weakens where the vertical's workflows become commodity integrations; which regime obtains in a given vertical is an empirical question the research agenda assigns to deployment-level observation.

Proposition 3 (Joins carry rarity). *Workflow data satisfies rarity in proportion to the workflow's cross-entity linkage.*

The variables are measured independently: linkage from the number and type of organisational boundaries the deployment's execution spans; rarity by documented survey of whether the vertical's data market sells or pools a functional substitute (Section 4.3). Rarity so assessed is a property of joins rather than volume: the constituent records of a boundary-spanning workflow may each be commonplace while their linkage exists nowhere else. The mechanism also imports Hagiu and Wright's (2023) across-customer learning condition into low-frequency settings — linkage aggregates sparse per-entity events into a corpus dense enough to learn from. The strongest objection is the non-rivalry premise (Jones and Tonetti, 2020): nothing in the nature of the joined dataset prevents its being sold or shared. The answer is institutional rather than physical: the joins arise as a by-product of a position of trust spanning the boundaries, and the confidentiality obligations attaching to that position are exactly what prevents their alienation — the same constraint that blocks resale blocks imitation, converting a legal restriction into an isolating mechanism (ex post limits to competition in Peteraf's (1993) terms).

The answer, however, cuts in both directions, and the proposition must carry the tension explicitly: the obligations that isolate the join also constrain the operator's own exploitation of it, since cross-entity generalisation from confidential material is permissible only where the position's consent architecture — aggregation rights, de-identification standards, purpose limitations, and the data-protection regime governing the records — has been deliberately constructed to permit it. Proposition 3 therefore holds only for operators whose boundary-spanning position is accompanied by lawful exploitation rights over the join; where the consent architecture is absent, the same mechanism that defeats rivals defeats the incumbent, and the join isolates no one. The scope condition is observable — consent and data-use provisions are documentary — and Design 2's contract analysis is specified to capture it. So stated, the proposition is refutable in both directions: high-linkage workflows whose deposits prove purchasable or pooled — the trajectory the vertical's own listing institutions exemplify (Section 6.1) — disconfirm it, as would single-perimeter workflows whose deposits turn out to have no market substitute.

Proposition 4 (Cadence compounds). *The value of workflow data — its capacity to improve execution of the generating and adjacent workflows — compounds with accumulation cadence, so that between two otherwise comparable positions the higher-cadence workflow yields the more durable advantage.*

Cadence governs the pace of the data feedback loop (Farboodi and Veldkamp, 2021) and, in Hagiu and Wright's (2023) terms, the steepness of the within-relationship learning curve; it further interacts with Proposition 2, since every period of a rival's co-adaptation effort widens the incumbent's lead. The strongest objection is the mediated-value finding: accumulation confers nothing absent the organisational capability to exploit it (Fosso Wamba et al., 2016; Mikalef, Pappas, Krogstie and Giannakos, 2017). The framework accepts the premise and locates its answer in the construct itself — in operated-mode provision the exploiting capability is the operator's own process design, so cadence and exploitation are not separable variables but joint properties of the same workflow; the objection therefore binds only in tool-mode configurations, where it reinforces Proposition 1.

Proposition 5 (Operator accrual). *In vendor-operated deployments, the workflow-data resource accrues by default to the operator rather than to the client whose activity generates it; contract can shift the asset's artefact components toward the client, but its tacit residue follows the operator under any drafting.*

The artefact components are those that can be specified and transferred — records, structured datasets, learned configurations, escrow of fine-tuned models — and their contractual assignment should shift measurable post-termination capability toward clients. The default follows from the deployment-mode dimension: execution passes through the operator's systems, which hold the by-product in the operator's structure (Section 4.2). The partial-contractibility claim rests on the incomplete-contracts premise — some components of the accumulated asset cannot be specified ex ante and therefore follow possession (Grossman and Hart, 1986; Hart and Moore, 1990) — joined to the outsourcing literature's account of what vendors accumulate (Levina and Ross, 2003; Lacity, Khan and Willcocks, 2009): unlike personnel-embodied expertise, a portion of the machine-executed asset is artefact — specifiable, transferable, escrowable — and to that portion contract reaches. The property-rights logic also yields an incentive corollary the classical treatment would predict: where clients anticipate that the deposit's non-contractible components follow the operator, they should rationally under-invest in the co-adaptation effort — data hygiene, process documentation, integration cooperation — that deepens the deployment, while operators should over-invest in operator-specific structure; the corollary is stated here as an additional testable implication rather than assumed, and Design 2 can observe it in the parties' relative integration investments. The proposition is therefore doubly falsifiable, and in observable terms. Against the default: operated engagements without artefact-reaching provisions in which clients nonetheless retain post-termination workflow performance — measured, per Design 2, as the client's performance under a successor operator or internal operation relative to pre-termination baseline — would refute the accrual claim. Against partial contractibility, from either side: if artefact-reaching clauses produce no measurable difference in client post-termination capability, contract reaches less than claimed; if well-drafted clauses eliminate the operator's advantage entirely, the tacit residue carries less than claimed and the distributional concern largely dissolves. The proposition's significance is distributional either way: the resource-based criteria may be satisfied by an asset the nominal beneficiary of the deployment does not own — a configuration the platform-derived literature does not raise because platform users are not buyers of operated services, and one on which the practical guidance of Section 8.2 directly depends.

Table 3 collects the propositions with their mechanisms, disconfirmers, and assigned designs; it is intended to make the framework's falsification structure visible at a glance. P1–P4 are comparative claims, each refutable by its non-monotonicity; P5 is independent of the first four — it concerns the allocation of whatever forms, not whether it forms. A field in which P1–P4 held but P5's default failed would vindicate the framework's economics while overturning its distributional conclusion. That configuration is the one most congenial to the vertical's incumbents, and therefore the one whose testing most requires independence from practitioner interest.

Table 3. Propositions: mechanism, disconfirming evidence, and testing design.

Proposition	Mechanism (source)	Disconfirmed by	Design (§7)
P1 Mode dominance	Generation, structure, and exploitation joined in one actor (Farboodi and Veldkamp, 2021)	Tool-mode telemetry jointly satisfying the criteria	1
P2 Time compression	Asset specificity expressed dynamically (Williamson, 1979; Teece, 1986)	Deep positions matched as fast as shallow upon displacement	1, 3
P3 Joins carry rarity	Boundary-spanning position + confidentiality as isolating mechanism (Peteraf, 1993)	High-linkage deposits proving purchasable or pooled; substitute-free single-perimeter deposits	2, 3
P4 Cadence compounds	Data feedback loop; within-customer learning (Farboodi and Veldkamp, 2021; Hagiu and Wright, 2023)	Value failing to compound with independently measured cadence	3
P5 Operator accrual	Residual control over non-contractible components (Grossman and Hart, 1986; Hart and Moore, 1990)	Clients retaining post-termination capability absent artefact provisions; artefact clauses producing no capability difference	2

Together the propositions restate the section's answer to the re-posed question. Workflow data constitutes a strategic resource where position is operated rather than supplied (P1), integration is deep enough to impose re-execution costs on imitators (P2), the workflow spans boundaries whose deposit the data market does not substitute (P3), and cadence compounds the position faster than rivals close it (P4); and where those conditions hold, the resource belongs to whoever operates the workflow, with contract able to shift its artefact components — though not its tacit residue — toward the client (P5). One clarification of the propositions' structure: P1's distinct content is the mode comparison (telemetry deposit against operated deposit); the monotonicity of criterion satisfaction across the remaining dimensions is carried by P2–P4 jointly, so that the framework's dimensional claims stand or fall as a conjunction rather than through P1 alone. The sceptical position survives as the special case at the tool-mode, shallow, low-linkage pole; the affirmative position survives as the operated-mode, deep, high-linkage case; and the disagreement between them, so restated, becomes a set of measurable contingencies rather than a contradiction.

6. Application: the real-estate vertical

The framework's value is best assessed against a concrete vertical, and real estate is selected here for a reason internal to the argument: its workflow characteristics place it close to the opposite pole, on every relevant dimension, from the consumer platforms on which the affirmative data-moat theory was built. If the framework applies where platform-derived mechanisms should be weakest, the unit-of-analysis shift has done genuine work.

6.1 Workflow characteristics of the vertical

Three characteristics of real-estate work, each documented in the vertical's own literature, map directly onto the framework's dimensions.

First, the vertical's core events are episodic and high-stakes rather than continuous. Transactions are infrequent per party, extended in duration, and mediated by chains of professional intermediaries — a workflow that Saull, Baum and Braesemann (2020) examine end-to-end, documenting both its length and the number of distinct parties whose sequential actions constitute it. In the framework's terms, per-entity accumulation cadence is low: no single client relationship generates the event volume on which within-customer learning of the platform kind (Hagiu and Wright, 2023) could operate. The framework accordingly predicts that durable positions in this vertical cannot be built on single-relationship learning and must instead arise, if at all, from Proposition 3's mechanism — linkage across entities that aggregates sparse events into learnable density.

Second, the work is document-intensive and its processes resist standardisation. The transaction workflow studied by Saull, Baum and Braesemann (2020) is composed substantially of document production, verification, and exchange among parties, and the literature on big data in corporate real-estate processes similarly locates the operational burden in heterogeneous, process-embedded records (Battisti, Shams, Sakka and Miglietta, 2019). Exception-rich, document-heavy processes are precisely the standardisation-resistant regime within which Proposition 2's scope condition is satisfied: co-adaptation to a particular organisation's documents, conventions, and exceptions is costly, and integration depth therefore carries its full imitability effect.

Third, the vertical's data is fragmented across organisational boundaries — but its history with that fragmentation is instructive in both directions, and the framework must be read against it. Braesemann and Baum (2020), asking whether real estate is becoming a data-driven market, document a landscape in which the relevant records are dispersed across brokerages, managers, lenders, and public registries, with proptech entrants positioning themselves precisely as partial aggregators; the absorptive-capacity findings of Vigren, Kadefors and Eriksson (2022) indicate limited internal digitalisation capability among incumbent owners, which renders the internal construction of such joins — a more demanding undertaking still — correspondingly unlikely. Fragmentation is the precondition of Proposition 3: where the joins do not otherwise exist and are not internally constructible, a boundary-spanning workflow generates a deposit for which the substitute survey of Section 4.3 finds no market equivalent.

The vertical's history, however, also supplies the proposition's boundary, and any application that ignored it would be incomplete. Real estate's most consequential data institutions are themselves boundary-spanning joins that were subsequently *pooled or commercialised* — through two distinct governance paths whose difference enriches rather than embarrasses the proposition. The multiple-listing service joins inventory records across competing brokerages as a member-gated club good — a pooling arrangement whose relative platform performance has been studied directly (Hendel, Nevo and Ortalo-Magné, 2009) — while the commercial data platforms documented in the proptech landscape (Braesemann and Baum, 2020) sell cross-entity records — listings, transactions, comparables — as products in open commerce. Pooling and sale extinguish rarity by different routes: the one by mutualising the join among its contributors, the other by pricing it for all comers; both illustrate the same terminal state for a standardisable join. These institutions are not counterexamples to Proposition 3; they are demonstrations of its mechanism running to completion. Where a join's value is high and its contents are standardisable, the vertical's actors have historically resolved the tension that non-rivalry creates (Jones and Tonetti, 2020) by institutionalising the join — pooling it mutually or selling it commercially — at which point rarity is extinguished and the sceptical predictions hold, exactly as the framework's tool-mode pole prescribes. The framework's prediction for the vertical is therefore discriminating rather than general: data advantage cannot form in the *already-pooled* joins (listing inventory, recorded transactions, comparables), and can form only in the residual — the operational and workflow-level joins that resist standardisation and pooling because they are exception-rich, confidential, and embedded in ongoing service relationships. The vertical's institutional history thus supplies both the baseline against which rarity must be assessed and a standing warning: any workflow join that standardises invites the multiple-listing outcome, and Proposition 3's rarity is always provisional on the join resisting institutionalisation.

6.2 What the framework predicts for the vertical

Against these characteristics, the framework yields three predictions, stated here as analytical implications rather than evidenced findings — the distinction the research agenda of Section 7 exists to close.

The first concerns *where* in the vertical data advantage can accumulate. Because per-entity cadence is low (Section 6.1), positions confined to a single client or transaction stream should yield little; positions that operate recurring workflows across many entities — portfolios, brokerage networks, multi-client service relationships — concentrate the vertical's sparse events into the density Propositions 3 and 4 require. The prediction is consistent with the observed strategy of proptech aggregators (Braesemann and Baum, 2020) but extends beyond platforms: any operated, boundary-spanning service position qualifies, at whatever scale.

The second concerns *who* is positioned to hold such advantage. The vertical's incumbents suffer documented absorptive-capacity constraints (Vigren, Kadefors and Eriksson, 2022), and its professional intermediaries have retained their role through digitalisation rather than being displaced by it (Asensio-Soto and Navarro-Astor, 2022; 2025). Proposition 5 then implies a specific and testable distributional claim: as artificial-intelligence systems enter the vertical's workflows in operated rather than tool mode, the resulting data resource will accrue to operators — vendors and intermediaries — rather than to the property-holding firms whose activity generates it, with contractual provision able to shift the asset's artefact components, though not its tacit residue, toward the client.

The third concerns the *limits* of data advantage in the vertical, and Section 6.1's institutional history has already stated the general form: advantage cannot form in joins that have been pooled or commercialised. Automated valuation is the canonical workflow-level instance — the vertical's longest-standing application of statistical and machine-learning methods (Abidoye and Chan, 2017), whose inputs and outputs alike circulate through the commercial data landscape (Braesemann and Baum, 2020) — and for it Proposition 3 denies rarity regardless of execution quality, predicting commodity competition rather than moat formation. The framework thus does not predict a general "AI land grab" in real estate: it partitions the vertical's workflows into those whose deposits are reconstructible from purchasable or pooled data, where the sceptical position should hold, and those whose deposits exist only as unpooled, confidential, boundary-spanning joins, where it should not.

A worked illustration in the journal's core domain fixes the predictions. Consider lease abstraction — the extraction of critical terms, dates, options, and covenants from heterogeneous lease documents into structured form, a standing burden of commercial property management and due diligence and an instance of the document-intensive process character documented for the vertical's workflows generally (Battisti, Shams, Sakka and Miglietta, 2019; Saull, Baum and Braesemann, 2020). Provided in tool mode — abstraction software licensed to a property manager — the framework predicts commodity competition: the vendor's telemetry sees documents-in and fields-out but not the exception resolution around them; the deposit is instrument-shaped; and comparable software is available to every rival at market prices. Provided in operated mode across a multi-client portfolio, every dimension shifts: the operator's process co-adapts to each client's lease conventions and exception cases (integration depth); the deposit joins abstracted terms across landlords, tenants, and jurisdictions that no single perimeter contains (cross-entity linkage); portfolio-wide throughput supplies the cadence that any single client's episodic transactions cannot; and the resulting corpus — covenant-level structures, option patterns, exception resolutions across landlords and jurisdictions — is the deposit on which Section 4.3's substitute survey would actually be run. Run informally, the survey illustrates the method's discipline: commercial lease-comparable products already pool and sell headline economics — rents, terms, concessions — across markets, so the corpus's substitutes are partial rather than absent, and its rarity resides in what the pooled products do not carry: exception resolution, covenant-level structure, and the coupling of both to an operating process. Whether that residual rarity survives is Proposition 3's live question in this workflow, and the §6.1 warning applies with full force: a standardised abstraction schema is precisely the kind of join the vertical has historically pooled.

6.3 The evidentiary gap as mismeasurement

To Section 3.3's finding — no identified study measures deployed-AI outcomes within real-estate firms — the framework adds a specific mismeasurement diagnosis: adoption and readiness instruments observe the tool-mode pole of the deployment-mode dimension, while Propositions 1 and 5 locate the competitively consequential variation at the operated pole, inside service relationships that firm-level surveys do not reach. The literature is thin in exactly the region where the framework predicts the phenomenon lives.

7. Research agenda

The propositions are stated to be testable, and the agenda that follows assigns each to the research design capable of testing it. The designs are ordered by evidentiary ambition; the first is deliberately within reach of a single investigator with field access, because the vertical's evidentiary base must be begun before it can be accumulated.

Design 1: Deployment-level multiple-case studies. Theory-building case research (Eisenhardt, 1989; Eisenhardt and Graebner, 2007; Yin, 2018) on artificial-intelligence deployments inside real-estate firms, with cases sampled to vary on the framework's dimensions — tool-mode against operated-mode provision, shallow against deep integration — and within-case evidence coded against the operationalised criteria of Table 2 using established qualitative procedure (Gioia, Corley and Hamilton, 2013). Such designs test Propositions 1 and 2 directly: they can observe whether by-product data accrues, to whom, in what structure, and whether performance in the generating workflow improves with accumulation. Because the natural investigators of operated deployments are often their operators, insider-research safeguards (Van de Ven and Johnson, 2006; Coghlan and Brannick, 2014) are a required element of the design rather than an optional refinement.

Design 2: Vendor–client dyad studies. Paired examination of both sides of operated deployments — the operator's accumulating asset and the client's post-deployment capability — including the content of data-rights clauses and the parties' respective positions at contract termination. The dyad is the natural unit for Proposition 5, whose distributional claim concerns precisely the asymmetry between the parties, and for the contractual-allocation question the platform literature cannot raise. Documentary method on the contracts themselves, joined to interviews on post-termination capability, would suffice; the design has no precedent in the vertical's literature, which has studied vendors (Tagliaro, Pomè, Migliore and Danivska, 2024) and adopting firms (Vigren, Kadefors and Eriksson, 2022) but never the dyad.

Design 3: Longitudinal accumulation measurement. Repeated observation of operated deployments over time, measuring accumulation cadence directly (events deposited per period), linkage (entities joined), and workflow performance (cycle time, exception rates) as they co-evolve. This is the design for Propositions 3 and 4, whose claims are inherently dynamic — rarity as joins accumulate, value as cadence compounds — and it supplies what cross-sectional adoption surveys cannot: the time path along which a position either compounds beyond imitation or fails to. The same design yields the vertical's first estimates of the imitability horizon that Proposition 2's re-execution logic presupposes.

Design 4: Cross-vertical comparison. Replication of Designs 1–3 in a second professional vertical differing on the framework's contextual variables — higher event cadence, lower document intensity, or standardised interfaces — to test the framework's own scope conditions, in particular the modularity boundary conceded to Clough and Wu (2020) under Proposition 2. Where workflows standardise, the framework predicts its own weakening; a comparison vertical in which integration depth is collapsing provides the natural experiment.

The designs are complementary rather than alternative, and their sequence is deliberate. Design 1 establishes existence and mechanism — whether the framework's configurations occur in the field and whether the criteria behave as operationalised. Design 2 isolates the allocation question, which no within-firm design can observe. Design 3 converts the framework's comparative statics into dynamics, and is the only design capable of estimating the imitability horizon — the re-execution period on which Proposition 2's entire logic rests — rather than assuming it. Design 4 disciplines the framework from outside, testing whether its scope conditions are doing real work or merely absorbing anomalies. An investigator beginning today would start with Design 1 not because it is easiest but because its output — coded, criterion-level deployment observations — is the input every subsequent design presupposes.

The agenda also implies instrumentation the vertical's research practice does not yet possess. Adoption and readiness surveys (Abidoye and Chan, 2017; Banaitienė and Cataldo, 2025) measure dispositions of firms toward technology; the framework's constructs are properties of *deployments* — mode, depth, linkage, cadence — and require instruments that take the deployment, not the firm, as the sampling unit. The nearest existing precedents are the workflow-level interview protocol of Saull, Baum and Braesemann (2020) and the ecosystem-level capability mapping of Vigren, Kadefors and Eriksson (2022); the agenda's designs extend the first to machine-executed workflows and the second across the vendor boundary. The shift in sampling unit is not a technicality: Section 6.3's mismeasurement argument implies that continued firm-level sampling will keep finding what it has always found — intentions, barriers, readiness — while the competitively consequential variation accumulates unobserved inside service relationships.

Two features of the agenda warrant emphasis. First, its constructs are already operationalised: Table 2 specifies what an investigator must observe, and the framework's dimensions are properties of deployments rather than of proprietary datasets, so the designs require field access but not data disclosure. Second, the agenda is falsifying as well as confirming: Design 1 can find that by-product data fails the value criterion even at the operated pole; Design 3 can find that accumulation does not compound; Design 4 can find that standardisation dissolves integration depth faster than positions form. The framework's contribution is to convert the data-moat controversy into observations worth making — in a vertical whose literature, as Section 3.3 established, has yet to make any of them.

8. Discussion and conclusion

8.1 Contribution

The paper set out from a debate the literature has left unfinished: the sceptical position on data-based advantage concludes with an open clause — transient *absent isolating mechanisms* — and the affirmative theory supplies candidate mechanisms specified in terms of learning structure but not of the organisational and institutional arrangements that instantiate them outside consumer platforms. The paper's contribution is a completion of that specification, and it is integrative by design: it does not claim novelty for the stock–process distinction, which the capability stream carries implicitly, nor for the underlying economics of specificity and residual control, which are due to transaction-cost and property-rights theory. Its claims are three.

First, the paper aligns the debate's open clause with the constructs capable of closing it. The workflow-ownership framework specifies which isolating mechanisms operate (co-specialisation, boundary-spanning joins, compounding cadence), where (along four observable dimensions of deployment), and in whose hands (the workflow's operator) — so that the positions of Lambrecht and Tucker (2015) and of Hagiu and Wright (2023) become jointly satisfiable descriptions of identified regions of one construct space, and the residual disagreement becomes a set of measurable contingencies rather than an unresolved dispute about the nature of data.

Second, the framework extends data-enabled-learning theory beyond the platform settings in which it was developed (C3), and it does so by joining that theory to literatures it has not previously met: the outsourcing stream's vendor-accumulation results (Levina and Ross, 2003; Lacity, Khan and Willcocks, 2009) and the incomplete-contracts analysis of residual control (Grossman and Hart, 1986; Hart and Moore, 1990). The junction is where the paper's most consequential question lives — what changes when the vendor's accumulating asset is partly a contractible artefact rather than personnel-embodied expertise — and it is a question none of the parent literatures poses alone.

Third, the paper supplies the firm-level analytical machinery for a claim so far argued only at industry level (C4): that advantage in the generative-artificial-intelligence era will rest on workflows and proprietary data rather than model access (Azoulay, Krieger and Nagaraj, 2024). The construct, the operationalised criteria, and the propositions convert that conjecture into objects of empirical study, and the research agenda assigns each proposition to a design capable of testing it in a vertical whose literature, so far as the present review could establish, contains no such observation.

A fourth implication, distributional rather than reconciliatory, may prove the most consequential for the vertical itself. Proposition 5 holds that the workflow-data resource accrues by default to whoever operates the workflow, and that the allocation is only partially contractible. If correct, the entry of operated artificial-intelligence services into real estate entails a reallocation of prospective resources from the firms whose activity generates the data to the vendors and intermediaries who operate the systems — a reallocation governed not by technology but by contract, currently observable principally in the drafting of data-rights provisions, and absent from the adoption literature's field of view.

8.2 Implications for practice

For real-estate operators, the analysis reframes the build-versus-buy decision. What is at stake in choosing operated provision is not only cost and capability but the location of an accumulating resource: a firm that buys operation cedes, by default, the deposit of its own activity. The practical instruments follow from Proposition 5's partial-contractibility claim — provisions reaching the asset's artefact components (structured datasets, learned configurations, escrow of fine-tuned models) shift measurable post-termination capability toward the client, while the tacit residue follows the operator under any drafting; the drafting question is therefore how much of the asset the contract can convert into artefact, not whether the contract "owns the data." For vendors, symmetrically, the analysis identifies which service positions compound (operated, deeply integrated, boundary-spanning, high-cadence) and which reduce to commodity competition however sophisticated the underlying models. For both parties, the framework implies that the widely discussed question of which artificial-intelligence system to adopt matters less than the largely undiscussed question of who operates it and on what data terms.

8.3 Limitations

The paper's limitations are those of its genre. It develops a framework and propositions from synthesis; it presents no new empirical evidence, and its predictions for the real-estate vertical in Section 6.2 are analytical implications awaiting the observation the agenda specifies. The review on which it rests is selective rather than exhaustive, and a portion of its corpus was engaged through verified metadata and abstracts rather than full text (Section 2). The framework's scope conditions are explicit but untested: where workflows standardise, integration depth may collapse faster than positions form (Clough and Wu, 2020), and which regime obtains in any given vertical is an empirical question. Finally, the propositions concern the conditions for a resource-based advantage, not its magnitude; nothing here establishes that workflow-data moats, where they form, are large.

8.4 Conclusion

Whether data confers durable advantage has been debated as a question about data. This paper has argued that it is better posed as a question about workflows: who executes the process that generates the data, how deeply that process is embedded in the organisations it serves, which boundaries it spans, and how fast its deposit compounds. So posed, the question admits determinate answers — five of them, stated as testable propositions — and a vertical, real estate, whose characteristics make it an unusually clean site for testing them. The data-moat debate need not be settled in the abstract; it can be settled in the field. The question of which system to adopt will be debated everywhere; the question of who operates it, and therefore who keeps what it learns, is the one this paper puts on the table.

References

Abidoye, R.B. and Chan, A.P.C. (2017), "Valuers' receptiveness to the application of artificial intelligence in property valuation", *Pacific Rim Property Research Journal*, doi: 10.1080/14445921.2017.1299453.

Asensio-Soto, J.C. and Navarro-Astor, E. (2022), "Proptech: A qualitative analysis of online real estate brokerage agencies in Spain", *Intangible Capital*, doi: 10.3926/ic.2090.

Asensio-Soto, J.C. and Navarro-Astor, E. (2025), "Relevance of the real estate agent in the Spanish residential market in the age of digitalisation", *International Journal of Strategic Property Management*, doi: 10.3846/ijspm.2025.25297.

Azoulay, P., Krieger, J. and Nagaraj, A. (2024), "Old Moats for New Models: Openness, Control, and Competition in Generative AI", working paper, National Bureau of Economic Research, doi: 10.3386/w32474.

Banaitienė, N. and Cataldo, I. (2025), "Human factors in PropTech adoption: Professional readiness and organisational challenges in Baltic real estate", *Human Technology*, doi: 10.14254/1795-6889.2025.21-3.6.

Barney, J.B. (1991), "Firm Resources and Sustained Competitive Advantage", *Journal of Management*, Vol. 17 No. 1, pp. 99-120, doi: 10.1177/014920639101700108.

Battisti, E., Shams, S.M.R., Sakka, G. and Miglietta, N. (2019), "Big data and risk management in business processes: implications for corporate real estate", *Business Process Management Journal*, doi: 10.1108/bpmj-03-2019-0125.

Braesemann, F. and Baum, A. (2020), "PropTech: Turning Real Estate Into a Data-Driven Market?", working paper, SSRN Electronic Journal, doi: 10.2139/ssrn.3607238.

Clough, D.R. and Wu, A. (2020), "Artificial Intelligence, Data-Driven Learning, and the Decentralized Structure of Platform Ecosystems", *Academy of Management Review*, doi: 10.5465/amr.2020.0222.

Coghlan, D. and Brannick, T. (2014), *Doing Action Research in Your Own Organization*, 4th ed., SAGE Publications, London, doi: 10.4135/9781529682861.

Eisenhardt, K.M. (1989), "Building Theories from Case Study Research", *Academy of Management Review*, Vol. 14 No. 4, pp. 532-550, doi: 10.5465/amr.1989.4308385.

Eisenhardt, K.M. and Graebner, M.E. (2007), "Theory Building from Cases: Opportunities and Challenges", *Academy of Management Journal*, Vol. 50 No. 1, pp. 25-32, doi: 10.5465/amj.2007.24160888.

Farboodi, M. and Veldkamp, L. (2021), "A Model of the Data Economy", working paper, National Bureau of Economic Research, doi: 10.3386/w28427.

Fosso Wamba, S., Gunasekaran, A., Akter, S., Ren, S.J.F., Dubey, R. and Childe, S.J. (2016), "Big data analytics and firm performance: Effects of dynamic capabilities", *Journal of Business Research*, doi: 10.1016/j.jbusres.2016.08.009.

Gioia, D.A., Corley, K.G. and Hamilton, A.L. (2013), "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology", *Organizational Research Methods*, Vol. 16 No. 1, pp. 15-31, doi: 10.1177/1094428112452151.

Goldfarb, A. and Tucker, C.E. (2019), "Digital Economics", *Journal of Economic Literature*, doi: 10.1257/jel.20171452.

Gregory, R.W., Henfridsson, O., Kaganer, E. and Kyriakou, H. (2020), "The Role of Artificial Intelligence and Data Network Effects for Creating User Value", *Academy of Management Review*, doi: 10.5465/amr.2019.0178.

Grossman, S.J. and Hart, O. (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration", *Journal of Political Economy*, doi: 10.1086/261404.

Gupta, M. and George, J.F. (2016), "Toward the development of a big data analytics capability", *Information & Management*, doi: 10.1016/j.im.2016.07.004.

Hagiu, A. and Wright, J. (2023), "Data-enabled learning, network effects, and competitive advantage", *The RAND Journal of Economics*, doi: 10.1111/1756-2171.12453.

Hart, O. and Moore, J. (1990), "Property Rights and the Nature of the Firm", *Journal of Political Economy*, doi: 10.1086/261729.

Hendel, I., Nevo, A. and Ortalo-Magné, F. (2009), "The Relative Performance of Real Estate Marketing Platforms: MLS versus FSBOMadison.com", *American Economic Review*, doi: 10.1257/aer.99.5.1878.

Jones, C.I. and Tonetti, C. (2020), "Nonrivalry and the Economics of Data", *American Economic Review*, doi: 10.1257/aer.20191330.

Krakowski, S., Luger, J. and Raisch, S. (2022), "Artificial intelligence and the changing sources of competitive advantage", *Strategic Management Journal*, doi: 10.1002/smj.3387.

Lacity, M.C., Khan, S.A. and Willcocks, L.P. (2009), "A review of the IT outsourcing literature: Insights for practice", *The Journal of Strategic Information Systems*, doi: 10.1016/j.jsis.2009.06.002.

Lambrecht, A. and Tucker, C.E. (2015), "Can Big Data Protect a Firm from Competition?", working paper, SSRN Electronic Journal, doi: 10.2139/ssrn.2705530.

Levina, N. and Ross, J.W. (2003), "From the Vendor's Perspective: Exploring the Value Proposition in Information Technology Outsourcing", *MIS Quarterly*, doi: 10.2307/30036537.

Mikalef, P., Pappas, I.O., Krogstie, J. and Giannakos, M.N. (2017), "Big data analytics capabilities: a systematic literature review and research agenda", *Information Systems and e-Business Management*, doi: 10.1007/s10257-017-0362-y.

Peteraf, M.A. (1993), "The Cornerstones of Competitive Advantage: A Resource-Based View", *Strategic Management Journal*, Vol. 14 No. 3, pp. 179-191, doi: 10.1002/smj.4250140303.

Saull, A., Baum, A. and Braesemann, F. (2020), "Can digital technologies speed up real estate transactions?", *Journal of Property Investment and Finance*, doi: 10.1108/jpif-09-2019-0131.

Souza, L.A., Koroleva, O.V., Worzala, E., Martin, C., Becker, A. and Derrick, N. (2020), "The technological impact on real estate investing: robots vs humans: new applications for organisational and portfolio strategies", *Journal of Property Investment and Finance*, doi: 10.1108/jpif-12-2020-0137.

Starr, C.W., Saginor, J. and Worzala, E. (2020), "The rise of PropTech: emerging industrial technologies and their impact on real estate", *Journal of Property Investment and Finance*, doi: 10.1108/jpif-08-2020-0090.

Tagliaro, C., Pome, A.P., Migliore, A. and Danivska, V. (2024), "Technology 'like a fork'. How PropTech shapes real estate innovation", *Journal of European Real Estate Research*, doi: 10.1108/jerer-05-2024-0035.

Teece, D.J. (1986), "Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy", *Research Policy*, Vol. 15 No. 6, pp. 285-305, doi: 10.1016/0048-7333(86)90027-2.

Teece, D.J. (2007), "Explicating Dynamic Capabilities: The Nature and Microfoundations of (Sustainable) Enterprise Performance", *Strategic Management Journal*, Vol. 28 No. 13, pp. 1319-1350, doi: 10.1002/smj.640.

Teece, D.J., Pisano, G.P. and Shuen, A. (1997), "Dynamic Capabilities and Strategic Management", *Strategic Management Journal*, Vol. 18 No. 7, pp. 509-533, doi: 10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.

Ullah, F., Sepasgozar, S.M.E. and Wang, C. (2018), "A Systematic Review of Smart Real Estate Technology: Drivers of, and Barriers to, the Use of Digital Disruptive Technologies and Online Platforms", *Sustainability*, doi: 10.3390/su10093142.

Van de Ven, A.H. and Johnson, P.E. (2006), "Knowledge for Theory and Practice", *Academy of Management Review*, Vol. 31 No. 4, pp. 802-821, doi: 10.5465/amr.2006.22527385.

Vigren, O., Kadefors, A. and Eriksson, K. (2022), "Digitalization, innovation capabilities and absorptive capacity in the Swedish real estate ecosystem", *Facilities*, doi: 10.1108/f-07-2020-0083.

Wernerfelt, B. (1984), "A Resource-Based View of the Firm", *Strategic Management Journal*, Vol. 5 No. 2, pp. 171-180, doi: 10.1002/smj.4250050207.

Williamson, O.E. (1979), "Transaction-Cost Economics: The Governance of Contractual Relations", *The Journal of Law and Economics*, doi: 10.1086/466942.

Yin, R.K. (2018), *Case Study Research and Applications: Design and Methods*, 6th ed., SAGE Publications, Thousand Oaks, CA.