Bottleneck-First Implementation: Applying Theory of Constraints to AI Adoption in Owner-Operated Small and Medium Businesses

Humberto Inciarte

May 2026

Abstract

Despite enterprise spending on generative artificial intelligence reaching an estimated USD 30–40 billion in 2024, 95% of organizations report no measurable profit-and-loss impact from their pilots, and the broader RAND Corporation analysis of AI projects finds a failure rate exceeding 80% — twice the rate of non-AI information-technology initiatives. This paper examines the dominant explanatory hypothesis embedded in the trade press — that artificial-intelligence adoption fails because of insufficient breadth — and argues, on the basis of converging evidence from three independent research traditions, that the prevailing diagnosis is structurally inverted. Drawing on Goldratt's Theory of Constraints (1984), Pareto-Juran prioritization theory, the Lean Production System taxonomy of *muda*, and contemporary empirical reports from RAND, the MIT Project NANDA initiative, the Boston Consulting Group, McKinsey & Company, and the OECD, this review synthesizes 20 sources across six thematic clusters. The analysis demonstrates that the documented failure pattern is mechanistically predicted by Theory of Constraints: optimization of non-bottleneck resources cannot increase system throughput. In owner-operated small and medium businesses, the binding constraint is empirically the owner, identified as the principal organizational risk in over 95% of middle-market assessments. The paper proposes a bottleneck-first implementation framework — operationalized as Agentes Para Tu Negocio — in which Goldratt's Five Focusing Steps are mapped onto the deployment of AI agents in owner-operated firms.

Affiliation: Department of Research, Arquitectos de la Felicidad / NeuroFlow 30H™

ORCID: 0009-0006-0379-7328 **Corresponding Author:** hola@humbertoinciarte.com

Keywords: Theory of Constraints, bottleneck analysis, SMB optimization, AI implementation strategy, business process priorities, Pareto principle, focused intervention, Agentes Para Tu Negocio, owner dependency, digital transformation

---

1. Introduction

1.1 The Prevailing Narrative

The dominant narrative surrounding artificial-intelligence adoption in small and medium-sized businesses (SMBs) has, since 2022, converged on a single prescription: comprehensive digital transformation. Trade press, vendor marketing, and consulting templates originally developed for the Fortune 500 have been progressively adapted downward, presenting SMB owners with a mandate to automate broadly across functions — sales, marketing, operations, customer service, finance — and, increasingly, simultaneously. This prescription is reinforced by aggregate investment data: the MIT Project NANDA initiative documents enterprise generative-AI investment of approximately USD 30–40 billion during 2024 (Challapally et al., 2025), and the Boston Consulting Group's 2024 cross-sectoral

survey of 1,000 chief executives finds that the average organization is pursuing dozens of parallel AI initiatives (Boston Consulting Group, 2024).

Two variants of this narrative dominate SMB-owner decision-making. The first is maximalist — "I need to automate my entire business" — and frames partial implementation as inadequate or transitional. The second is procedural — "I will start with the easiest thing" — and frames implementation sequencing as a matter of operator convenience rather than strategic impact. Both variants share an implicit assumption: that the value of AI adoption scales primarily with breadth of application.

1.2 The Problem

The empirical record contradicts this assumption. The RAND Corporation, in a 2024 study based on structured interviews with 65 senior data scientists and machine-learning engineers, reports that more than 80% of AI projects fail — approximately twice the failure rate of non-AI information-technology initiatives, with the principal root cause identified as "misunderstanding the business problem" (Ryseff & Narayanan, 2024). The MIT Project NANDA *State of AI in Business 2025* report, surveying enterprise generative-AI deployments, finds that 95% of pilots produce no measurable profit-and-loss impact, despite the large aggregate investment cited above (Challapally et al., 2025). The Boston Consulting Group reports that only 4% of companies generate substantial value from AI at scale, with an additional 22% characterized as emerging leaders and 74% reporting no tangible value (Boston Consulting Group, 2024). An IDC study referenced by *CIO* magazine indicates that 88% of AI proofs-of-concept never reach production deployment (Bednarz, 2024).

These data points share a structural feature that is rarely commented upon in the trade press: they describe failure at scale that is uncorrelated with technology selection, model quality, or vendor choice. The failure pattern persists across model generations, across deployment contexts, and across firm sizes. As Davenport and Ronanki (2018) documented during the previous wave of cognitive-computing adoption, "companies do better by taking an incremental rather than a transformative approach" — a finding that pre-dates the current generative-AI cycle by half a decade and suggests the failure pattern is not technology-specific but methodologically structural.

1.3 Research Question and Thesis

This paper examines a single research question: *why does AI adoption fail at such high rates in owner-operated SMBs, and what implementation sequence improves outcomes?* The thesis advanced is that the dominant failure mode is not technological but structural — that AI is being deployed away from the system's binding constraint, which the Theory of Constraints (Goldratt, 1984) predicts will produce zero net throughput improvement regardless of the technology's intrinsic capability. In owner-operated SMBs, the binding constraint is empirically the owner. A bottleneck-first implementation methodology, in which a single AI system is designed to act on the identified constraint before any breadth is pursued, is therefore proposed as a necessary starting condition for SMB AI return-on-investment.

2. Literature Review

2.0 Methodological Note

This review synthesizes peer-reviewed empirical studies, organizational case reports, and institutional data published between 1984 and 2026, sourced from Google Scholar, ResearchGate, Semantic Scholar, JSTOR, SSRN, and the publication archives of established research organizations including the RAND Corporation, the MIT Initiative on the Digital Economy, the Boston Consulting Group, McKinsey & Company, the OECD, and the International Data Corporation. Inclusion criteria prioritized works with measurable outcomes related to (a) constraint-based operations theory, (b) AI or digital-transformation adoption results, or (c) owner-dependency in privately held firms. Where peer-reviewed evidence was limited — particularly in cluster 2.4 (owner dependency) — proprietary industry data from Class VI Partners, Pinnacle Equity Solutions, the Exit Planning Institute, and Permanent Equity were included with explicit identification as institutional rather than peer-reviewed sources. The total source base comprises 20 references organized into six thematic clusters.

2.1 Theory of Constraints — The Foundational Operations Framework

The Theory of Constraints was introduced in monograph form by Eliyahu Goldratt in *The Goal: A Process of Ongoing Improvement* (Goldratt & Cox, 1984), published by North River Press. The theory's central proposition is that every system, regardless of complexity, contains exactly one binding constraint — a single resource, policy, or market condition that determines the system's maximum throughput. Improvements made to non-constraints do not increase system throughput; they only increase local efficiency at the cost of system-level imbalance. This proposition is operationalized through what Goldratt subsequently termed the Five Focusing Steps: identify the system's constraint, decide how to exploit it, subordinate every other resource to that decision, elevate the constraint, and — once a constraint is broken — return to step one to avoid inertia (Goldratt, 1990).

The theoretical legitimacy of TOC was formally evaluated by Naor, Bernardes, and Coman (2013), who applied the eight Popper-Wacker criteria for theory validity (generalizability, parsimony, fecundity, internal consistency, empirical riskiness, abstraction, conservation, and conceptual coherence) and concluded that TOC satisfies each, situating it as a legitimate operations-management theory rather than a consultant heuristic. Empirical validation comes principally from Mabin and Balderstone (2003), whose meta-analytic review synthesized 80 documented TOC implementations: mean reported outcomes were a 70% reduction in lead time, a 65% reduction in cycle time, a 49% reduction in inventory, and a 63% increase in revenue. The authors note, with appropriate methodological caution, that "despite extensive searches, the research found no reports of failures" — a finding that simultaneously suggests TOC's robustness and signals likely publication bias.

The extension of TOC beyond its manufacturing origins is documented in the *Theory of Constraints Handbook* (Cox & Schleier, 2010), covering applications in services, healthcare, project management, and information technology. Goldratt himself, in later writings, distinguished three constraint types — physical, market, and policy — and observed that real-world firms rarely face market constraints; instead, they typically face policy or behavioral constraints embedded in their own operating decisions

(Goldratt, 1990). This distinction is central to the present paper's application of TOC to SMBs, in which the binding constraint is rarely physical capacity but rather a policy or behavioral structure embedded in owner dependence.

2.2 Pareto and Lean — Independent Confirmation of the Focused-Intervention Asymmetry

A second research tradition arrives at a structurally similar prescription via different reasoning. The Pareto principle, originally a 19th-century statistical observation regarding land ownership in Italy, was generalized to operations management by Joseph Juran in his *Quality Control Handbook* (Juran, 1951), where it became the principle of the "vital few and the useful many" — the empirical observation that approximately 80% of quality outcomes derive from approximately 20% of root causes. Juran himself emphasized in later writings that prioritization without identification of the vital few produces effort without effect.

The Toyota Production System, formalized by Taiichi Ohno (1988) and codified for Western audiences by Womack and Jones (1996) in *Lean Thinking*, classifies seven categories of waste, or *muda*: overproduction, waiting, transport, over-processing, inventory, motion, and defects. Of these, two — over-processing and overproduction — are explicitly defined as activities that occur outside the constraint and create the appearance of progress while consuming resources without increasing system output. Womack and Jones argue that lean's value-stream mapping discipline exists precisely to prevent the optimization of non-constraint activities, which they describe as "an illusion of progress."

The convergence is significant. TOC, derived from queuing theory and operations research, prescribes constraint-first improvement on grounds of system throughput. Lean, derived from Toyota's empirical practice, prescribes constraint-first improvement on grounds of waste elimination. Pareto-Juran, derived from quality-control statistics, prescribes constraint-first improvement on grounds of distributional asymmetry. Three independent research traditions, developed in different decades and intellectual contexts, arrive at the same prescription.

2.3 The Empirical Failure of Broad AI Adoption

The AI-adoption literature, drawn principally from 2018–2026 institutional research, documents a failure pattern consistent with the predictions of clusters 2.1 and 2.2. Davenport and Ronanki (2018), in a *Harvard Business Review* analysis of 152 cognitive-AI projects across 250 surveyed executives, document the case of MD Anderson Cancer Center's USD 62 million Watson deployment — a moonshot transformation eventually shelved without clinical use, while narrower IT-cognitive deployments at the same institution succeeded. The authors conclude that "companies do better by taking an incremental rather than a transformative approach."

The RAND Corporation study (Ryseff & Narayanan, 2024) identifies five root causes of AI project failure, of which the foremost is "industry stakeholders often misunderstand — or miscommunicate — what problem needs to be solved using AI." The remaining four — inadequate data, insufficient infrastructure, talent gaps, and chasing the wrong technology — describe, in TOC vocabulary, attempts

to subordinate non-constraint resources before the constraint has been correctly identified.

The MIT Project NANDA *State of AI in Business 2025* report (Challapally et al., 2025) provides what is, at present, the most quantitatively granular evidence in this domain. Vendor-led, narrowly scoped implementations succeed at approximately 67%, while broad internal builds succeed at approximately 33% — a roughly 2× ratio favoring the focused approach. Mid-market firms scale successful pilots in approximately 90 days, against approximately 9 months in large enterprises, suggesting that organizational distance between constraint identification and constraint action is itself a determinant of success — a finding compatible with Goldratt's emphasis on subordination latency.

Boston Consulting Group's *Where's the Value in AI?* analysis (Boston Consulting Group, 2024) reports that AI leaders pursue approximately half as many initiatives as laggards but expect more than twice the return on investment, allocating resources in a 10/20/70 ratio across algorithms, technology, and people-and-process. An IDC analysis cited by Bednarz (2024) reports that 88% of AI proofs-of-concept never reach production. The OECD (2025) finds firm-level integration of AI into core business functions ranges from 1.9% in Japan to 6.1% in the United States, even where individual generative-AI use rates exceed 40% — confirming that the binding constraint is not access to tools but methodology of deployment.

A particularly consequential finding emerges from research on autonomous AI agents. A Stanford University and Carnegie Mellon University study reports that full automation of complex workflows produced a 17.7% performance *decrease*, while targeted augmentation of constrained activities produced a 24.3% performance *increase* (cited in Boston Consulting Group, 2025). The asymmetry is consistent with Goldratt's stronger formulation: an hour saved at a non-bottleneck is not merely worthless but can be net-negative when it produces unabsorbable capacity that destabilizes the constrained step.

2.4 The Owner-Dependency Phenomenon in Privately Held Firms

The locus of the binding constraint in SMBs is documented in a substantial industry-research literature. Class VI Partners' proprietary CoPilot assessment data identify "business too dependent on the owner" as the principal organizational risk in more than 95% of middle-market assessments — a prevalence cited by the firm as approximately 14 percentage points higher than the next most common risk (Class VI Partners, 2023). The Pinnacle Equity Solutions Owner Dependence Index, applied across a sample of 560 operating-company owners, reports a national mean dependence score of 52% — that is, the median privately held company is more than half dependent on the individual owner's daily efforts (Maddox, 2016).

The Exit Planning Institute's *2023 National State of Owner Readiness Report* documents that 80% of private-business owners' net worth is concentrated in their company, that 75% intend to exit within ten years, and that 49% have no transition plan; only 20–30% of businesses listed for sale ultimately close, with owner-dependency repeatedly identified as a top deal-killer (Exit Planning Institute, 2023). Permanent Equity's 2022 analysis quantifies the financial cost: each additional degree of founder control reduces enterprise value by between 23.0% and 58.1% (Permanent Equity, 2022).

These findings — drawn from independent assessment methodologies, samples ranging from 560 to thousands of firms, and industry researchers with no shared institutional affiliation — converge on a single empirical claim: in privately held firms, the owner is, in operations-theoretic terms, the binding constraint. This conclusion holds across firm size, industry, and geography to a degree unusual in business-research literatures.

2.5 Operational Excellence and the McKinsey Bottleneck Vocabulary

McKinsey & Company's *Today's Good to Great: Next-Generation Operational Excellence* report (McKinsey, 2024) reports that only 12% of organizational transformation programs sustain performance gains beyond three years, and explicitly cautions that "deploying technology without understanding the real sources of an issue is a recipe to make matters worse rather than better. Automating a low-productivity assembly line can simply speed up poor quality." The same analysis describes financial-services cases in which 90% of IT capacity was consumed by support work, leaving only 10% for value creation; targeted bottleneck intervention reduced application-operations spend by more than 30% and improved stability by more than 20% within six months (McKinsey, 2024). These findings are notable both for the explicit use of "bottleneck" vocabulary in a top-tier consultancy publication and for the documented magnitude of focused-intervention returns.

2.6 Research Gap

The literature reviewed in clusters 2.1 through 2.5 establishes four propositions independently: (a) Theory of Constraints provides a validated, falsifiable framework for predicting which interventions in a system will increase throughput; (b) Pareto-Juran and Lean traditions independently confirm the asymmetric returns of focused interventions; (c) AI adoption in 2018–2026 fails at rates between 80% and 95% in patterns consistent with TOC's predictions; (d) owner dependency is the empirically dominant constraint in privately held SMBs. Despite the proximity of these literatures, no source located in this review explicitly synthesizes them to derive an implementation methodology. The present paper proposes such a synthesis.

3. Analysis and Discussion

3.1 The Convergence Argument

The principal analytical contribution of this review is the observation that three independent research traditions, developed across approximately a century of operations-management thought, predict the same outcome: focused interventions targeting the system's binding constraint outperform broad interventions by a factor that the literature situates between approximately 2× and 3×. The 2× estimate is anchored in MIT Project NANDA's 67%-versus-33% success-rate differential (Challapally et al., 2025) and in the Boston Consulting Group's "leaders pursue half as many initiatives, expect twice the return" finding (Boston Consulting Group, 2024). The Mabin and Balderstone (2003) TOC meta-analysis reports outcome magnitudes — 70% lead-time reduction, 49% inventory reduction, 63%

revenue increase — that are difficult to translate into a single ratio but consistent with focused interventions dominating broad ones by a similar order of magnitude.

The Stanford-CMU agent-workflow finding, in which full automation produced a 17.7% performance decrease while targeted augmentation produced a 24.3% increase (cited in Boston Consulting Group, 2025), provides what is, in this review, the most consequential single empirical anchor. It demonstrates that the asymmetry is not merely between focused-and-superior and broad-and-inferior outcomes; rather, broad outcomes can be net-negative. This is consistent with Goldratt's stronger formulation in *The Goal*: an hour saved at a non-bottleneck is not merely useless but actively destabilizing, because it creates capacity that the constrained step cannot absorb, generating queue distortion and reducing system-wide reliability.

3.2 The Owner as the Binding Constraint — Structural Analysis

The application of Theory of Constraints to owner-operated SMBs yields a specific structural diagnosis. Class VI Partners' identification of owner-dependency as the principal risk in over 95% of middle-market assessments (Class VI Partners, 2023), combined with Pinnacle Equity Solutions' median Owner Dependence Index of 52% (Maddox, 2016) and Permanent Equity's documentation that each degree of founder control reduces enterprise value by 23–58% (Permanent Equity, 2022), establishes the owner as the binding constraint with empirical force comparable to manufacturing-floor bottleneck identification in classical TOC case studies.

The mechanism is well documented in the practitioner literature: SMBs accumulate operational knowledge tacitly in the owner's daily routines — pricing judgment, customer-relationship history, vendor negotiation patterns, quality-control intuition. Documentation and delegation infrastructure are typically absent or partial. The result is a system in which the owner's working hours constitute the literal bottleneck of throughput: orders accepted, quotes prepared, customer issues resolved, and strategic decisions made all queue at the owner's calendar.

In the TIMER vocabulary used in operational diagnostics — Time, Money, and Energy as separable currencies — the costs are quantifiable. *Time*: surveys consistently show owner-operators reporting 50- to 60-hour working weeks with no relief mechanism. *Money*: an owner who allocates ten weekly hours to undirected experimentation with AI tools, valued conservatively at USD 100 per hour, expends USD 4,000 monthly without diagnostic anchoring; over four months this represents a USD 16,000 opportunity cost with no documented system implementation. *Energy*: decision-quality decay under sustained cognitive load is well-documented in the cognitive-fatigue literature, suggesting that sustained owner-as-bottleneck operation degrades not only output volume but output quality.

3.3 Why Comprehensive Transformation Fails — A Lean/TOC Diagnosis

The empirical failure rates documented in cluster 2.3 are mechanistically explained by the joint application of Lean and TOC vocabularies. A comprehensive AI transformation initiative in an owner-bottlenecked SMB simultaneously incurs three categories of waste: *over-processing*, in which the automation of activities that do not lie on the constraint produces local efficiency without system

throughput change (a customer-service chatbot installed while the owner remains the bottleneck for sales-quote generation reduces one queue while leaving the binding queue untouched); *overproduction*, in which capacity created at non-constraint stations cannot be absorbed by the constrained step (marketing-automation that doubles inbound lead volume into an owner-bottlenecked sales process produces lead decay rather than revenue); and *local-optimum subordination failure*, in which the natural tendency of comprehensive initiatives is to optimize each function independently, violating Goldratt's third focusing step.

This explains, in TOC terms, why MIT NANDA finds vendor-led narrow implementations succeeding at 67% while broad internal builds succeed at 33%: the vendor-led approach is structurally incapable of broad simultaneity, and the narrow scope forces — by accident of delivery model — alignment with the focusing-steps discipline. The failure of broad internal builds is not a failure of capability but a failure of method.

A compounding factor is cognitive. The "law of the instrument" — sometimes called Maslow's hammer — describes the tendency to apply newly available tools indiscriminately. The arrival of generative AI in 2022–2023 produced precisely this pattern at scale: SMB owners encountered a powerful new tool and applied it to whatever activities were proximate to their attention, rather than to whatever activity was structurally constrained. The result is what one might describe as productivity theater: visible activity, measurable engagement, no throughput change.

3.4 Proposed Framework: Bottleneck-First Implementation as Operationalized TOC

The framework proposed in this paper — operationalized in practice as Agentes Para Tu Negocio — is presented not as a novel theoretical contribution but as a literal mapping of Goldratt's Five Focusing Steps onto the deployment of AI agents in owner-operated SMBs. Each component corresponds to one of Goldratt's steps:

Step 1 (Identify the constraint). A diagnostic phase precedes any system construction. The binding constraint is identified through structured analysis of the owner's working week, decision queues, and revenue-loss patterns, typically locating the constraint in one of four areas: sales-and-follow-up infrastructure, criterion-based evaluation (proposals, content, hiring decisions), repetitive operational responses, or method-and-delivery for knowledge businesses. This step requires criterion external to any specific tool, because a constraint is defined by its effect on system throughput — which the SMB owner often cannot see from inside the system.

Step 2 (Exploit the constraint). A single AI system, denoted the *Primer Sistema Estratégico* (First Strategic System), is designed to act on the identified constraint at maximum effect. The deliverable is one functioning system, not a roadmap and not a portfolio of pilots. The discipline is restrictive: not "what could AI do for this business" but "what must it do to relieve this specific constraint."

Step 3 (Subordinate everything else to the decision). Other automation, tooling, or process-improvement initiatives are deferred until the constraint-relief system is operational. This subordination is the most counter-cultural element of the framework, because the prevailing narrative encourages parallel pursuit. TOC predicts, and the AI-failure data confirm, that parallel pursuit

before constraint relief produces the documented 80–95% failure rate.

Step 4 (Elevate the constraint). Once the first system is operational, the owner's effective capacity in the constrained domain expands. This is not merely time-saving; it is structural relief, in that the system absorbs decisions or actions that previously queued at the owner. Elevation creates the conditions under which a new binding constraint emerges — typically further upstream or downstream — which becomes the target of the next intervention.

Step 5 (Avoid inertia). The framework explicitly cycles. After the first constraint is broken, the system is re-diagnosed; the prior solution is not assumed to still be optimal in the new configuration. This is consistent with Goldratt's caution against the most common implementation failure of TOC programs, in which an initial success becomes ossified into doctrine.

The framework's distinguishing features can be stated negatively: it does not pursue comprehensive transformation; it does not begin with the activity most convenient to the owner; it does not offer a general-purpose AI capability; and it does not present the technology as the deliverable. The deliverable is constraint relief, codified in a working system, supported by the diagnostic criterion that produced it. AI is the implementation vehicle, not the product.

3.5 Practical Implications

The findings have three orders of practical implication. For SMB owners, the principal implication is that the question "what AI tool should I adopt?" is structurally inverted; the prior question is "what is my system's binding constraint?" — answerable only through diagnosis, and rarely answerable from inside the system without external criterion. Beginning with tools and then seeking applications inverts the focusing-steps sequence and predicts the failure rates documented in cluster 2.3.

For implementers and consultants, the implication is that engagements promising broad transformation should be regarded with the skepticism warranted by their failure rate. The MIT NANDA finding that vendor-led narrow implementations succeed at twice the rate of broad internal builds is a methodological vindication of restrictive scope: practitioners who structure engagements around a single, constraint-targeted system — *one system before the next* — are, in TOC terms, practicing subordination correctly.

For vendors and platform providers, the implication is that product strategy emphasizing breadth of feature coverage may be optimizing the wrong objective. The Boston Consulting Group's 10/20/70 resource-allocation finding indicates that organizational and methodological factors dominate capability factors in determining outcomes. Platforms that include diagnostic and methodological scaffolding alongside technical capability are likely to outperform platforms that compete on capability alone.

A summary heuristic, useful as a methodological anchor: *implementation begins at the bottleneck, not at the tool*.

4. Conclusions

4.1 Summary of Findings

This review has documented four convergent propositions. First, the Theory of Constraints, originally introduced in Goldratt's 1984 monograph, provides a validated and falsifiable framework for predicting which interventions in a system will increase throughput; the framework satisfies the formal criteria of a legitimate operations theory (Naor et al., 2013) and has been empirically validated across more than 80 documented implementations (Mabin & Balderstone, 2003). Second, the Pareto-Juran and Lean research traditions independently confirm the asymmetric returns of focused-over-broad interventions, providing theoretical convergence from distinct intellectual histories. Third, the 2018–2026 empirical record of AI adoption — RAND's 80% failure rate, MIT NANDA's 95% pilot ineffectiveness, BCG's 4% substantial-value-at-scale figure, IDC's 88% production-deployment failure — describes a failure pattern consistent with TOC's prediction that non-constraint optimization cannot increase system throughput. Fourth, in owner-operated SMBs, the binding constraint is empirically the owner; this finding holds across firm size, industry, and assessment methodology with unusual consistency.

The synthesis yields the paper's central conclusion: comprehensive AI adoption in owner-operated SMBs is not merely sub-optimal but structurally predicted to fail at the rates observed in the empirical literature, and the corrective is not better technology, more data, or improved talent but a methodological reordering of implementation sequence to prioritize the binding constraint. Bottleneck-first implementation, operationalized through a literal application of Goldratt's Five Focusing Steps to AI-agent deployment, is therefore proposed as a necessary starting condition for SMB AI return-on-investment.

A concise restatement of the central insight, useful for practitioner communication: *AI without business judgment is, structurally, an industrial drill operating without knowledge of where the screw must be placed.*

4.2 Limitations

This study is subject to several limitations that warrant explicit acknowledgment. As a narrative literature review, the analysis is subject to the selection bias inherent in non-systematic synthesis; a formal systematic review with pre-registered search protocols would strengthen the evidentiary base. Several of the most consequential industry data points — particularly the Class VI Partners CoPilot assessment prevalence and the Pinnacle Equity Solutions Owner Dependence Index — derive from proprietary methodologies that have not been peer-reviewed, although their convergence with EPI and Permanent Equity findings strengthens triangulation. The Mabin and Balderstone (2003) meta-analysis, while comprehensive, is subject to the publication-bias caution the authors themselves raise: failures of TOC implementation are rarely published, and the absence of negative cases in their corpus may reflect reporting practice rather than method robustness. The framework proposed in section 3.4 has not been validated through a controlled experimental study; the paper's claim is that bottleneck-first implementation is a necessary, not necessarily sufficient, condition for SMB AI return. Finally, the AI-adoption empirics cited in cluster 2.3 are 2024–2026 vintage and will continue to evolve; the

paper's argument depends on the direction and order of magnitude of the failure rates rather than on specific decimal values.

4.3 Future Research Directions

Three lines of investigation appear particularly warranted by the analysis. First, a controlled or quasi-experimental study comparing bottleneck-first against breadth-first AI implementation in matched SMBs would address the principal evidentiary gap identified in section 4.2; the natural design would assign a sample of SMBs randomly to constraint-diagnostic-then-implement and to breadth-then-narrow protocols, with throughput, revenue, and owner-time-recapture as outcome measures over a six- to twelve-month window. Second, longitudinal validation of the Owner Dependence Index against post-AI-adoption outcomes would clarify whether dependency reduction is a leading indicator of return-on-investment, allowing diagnostic instruments to predict implementation success. Third, regional adaptation of the framework to Latin American SMB contexts — where the *Agentes Para Tu Negocio* methodology is currently operationalized — would address the present absence of Spanish-language empirical work in this domain and contribute to the cross-cultural generalizability of TOC-based AI adoption methodology.

Resumen en Español

A pesar de que el gasto empresarial en inteligencia artificial generativa alcanzó aproximadamente 30 a 40 mil millones de dólares en 2024, el 95% de las organizaciones reporta cero impacto medible en sus estados de resultados, y el análisis más amplio de la Corporación RAND sobre proyectos de IA encuentra una tasa de fracaso superior al 80%, el doble que la de iniciativas no-IA en tecnología de la información. Este artículo examina la hipótesis explicativa dominante en la prensa especializada — que la adopción de IA fracasa por insuficiencia de amplitud — y argumenta, sobre la base de evidencia convergente de tres tradiciones de investigación independientes, que el diagnóstico predominante está estructuralmente invertido. Sintetizando 20 fuentes de la Teoría de Restricciones de Goldratt (1984), la teoría de priorización Pareto-Juran, la taxonomía Lean del *muda*, y los reportes empíricos contemporáneos de RAND, MIT Project NANDA, Boston Consulting Group, McKinsey, OECD e IDC, el análisis demuestra que el patrón documentado de fracaso es predicho mecánicamente por la Teoría de Restricciones: la optimización de recursos que no son cuello de botella no puede aumentar el throughput del sistema. En PYMEs operadas por sus dueños, la restricción vinculante es empíricamente el dueño mismo, identificada como el principal riesgo organizacional en más del 95% de las evaluaciones del mercado medio. El artículo propone un marco de implementación bottleneck-first — operacionalizado como Agentes Para Tu Negocio — en el cual los Cinco Pasos de Enfoque de Goldratt se mapean sobre el despliegue de agentes de IA en empresas operadas por sus dueños.

Palabras clave: Teoría de Restricciones, análisis de cuello de botella, optimización de PYMEs, estrategia de implementación de IA, prioridades de procesos de negocio, principio de Pareto, intervención focalizada, Agentes Para Tu Negocio, dependencia del dueño, transformación digital

References

Bednarz, A. (2024, March 12). 88% of AI pilots fail to reach production — but that's not all on IT. *CIO*. https://www.cio.com/article/3850763/88-of-ai-pilots-fail-to-reach-production-but-thats-not-all-on-it.html

Boston Consulting Group. (2024). *Where's the value in AI?* https://www.bcg.com/publications/2024/wheres-value-in-ai

Boston Consulting Group. (2025). *AI leaders outpace laggards with double the revenue growth and 40% more cost savings*. BCG Press Release, September 30, 2025. https://www.bcg.com/press/30september2025-ai-leaders-outpace-laggards-revenue-growth-cost-savings

Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025). *The GenAI divide: State of AI in business 2025*. MIT Project NANDA. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Class VI Partners. (2023). *Business owner dependence: The risk hidden in most middle-market businesses*. CoPilot Assessment Report. https://www.classvipartners.com/business-owner-dependence-the-risk-hidden-in-most-middle-market-businesses/

Cox, J. F. III, & Schleier, J. G. (2010). *Theory of constraints handbook*. McGraw-Hill. ISBN 978-0071665544.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108–116. https://hbr.org/2018/01/artificial-intelligence-for-the-real-world

Exit Planning Institute. (2023). *2023 national state of owner readiness report*. Exit Planning Institute. https://exit-planning-institute.org/state-of-owner-readiness

Goldratt, E. M. (1990). *What is this thing called the theory of constraints and how should it be implemented?* North River Press.

Goldratt, E. M., & Cox, J. (1984). *The goal: A process of ongoing improvement*. North River Press. ISBN 978-0884271956.

Juran, J. M. (1951). *Quality control handbook* (1st ed.). McGraw-Hill.

Mabin, V. J., & Balderstone, S. J. (2003). The performance of the theory of constraints methodology: Analysis and discussion of successful TOC applications. *International Journal of Operations & Production Management*, 23(6), 568–595. https://doi.org/10.1108/01443570310476636

Maddox, G. (2016, July). *Recent business owner survey results: Companies are 52% dependent upon owners, insights behind the numbers*. Pinnacle Equity Solutions / LinkedIn Pulse. https://www.linkedin.com/pulse/recent-business-owner-survey-results-companies-52-greg-maddox-

McKinsey & Company. (2024). *Flip the ratio: Taking IT from bottleneck to battle ready*. McKinsey Digital. https://www.mckinsey.com/capabilities/operations/our-insights/todays-good-to-great-next-generation-operational-excellence

Naor, M., Bernardes, E. S., & Coman, A. (2013). Theory of constraints: Is it a theory and a good one? *International Journal of Production Research*, 51(2), 542–554. https://doi.org/10.1080/00207543.2011.654137

OECD. (2025). *AI adoption by small and medium-sized enterprises*. OECD Publishing.

Ohno, T. (1988). *Toyota production system: Beyond large-scale production*. Productivity Press.

Permanent Equity. (2022). *The kingdom or the crown: Addressing the owner dependence dilemma in your company*. https://www.permanentequity.com/content/the-kingdom-or-the-crown-addressing-the-owner-dependence-dilemma-in-your-company

Ryseff, J., & Narayanan, A. (2024). *Why AI projects fail and how they can succeed* (Report No. RR-A2680-1). RAND Corporation. https://www.rand.org/pubs/presentations/PTA2680-1.html

Womack, J. P., & Jones, D. T. (1996). *Lean thinking: Banish waste and create wealth in your corporation*. Free Press / Simon & Schuster.