

THE CAPACITY TO PROVE: Governance and Trust in the Age of AI

Introduction

Artificial intelligence (AI) allows firms to decide, communicate, contract, market, and extend credit at unprecedented speed while making responsibility and authenticity harder to establish. Synthetic content can resemble evidence; automated decisions can obscure accountability; environmental claims can outrun substantiation; platforms can restrict access to business data; code can execute obligations before disputes emerge; and frictionless credit can conceal consumer risk. These problems appear legally distinct but present a common governance challenge: can the firm prove that a consequential digital action was accountable, authentic, substantiated, and contestable?

This Policy Brief calls a digital action consequential when it can materially affect rights, money, contractual obligations, customer access, regulatory compliance, or reputation. A material AI application is an AI use that can make or materially influence such an action. Trust-by-design means integrating accountability, authenticity, substantiation, and contestability into the workflow before deployment or publication, rather than adding controls only after a failure. The result sought is verifiable trust: the firm's capacity to demonstrate how a consequential digital action occurred, who was responsible, what evidence supported it, and how an error can be challenged or corrected.

Written by:

Dr. Severo C. Madrona Jr.
Department of Commercial Law
RVR College of Business
DLSU-Manila

The brief proceeds through doctrinal analysis of Philippine electronic commerce and evidence law, a comparison with European Union, United States, and United Nations Commission on International Trade Law (UNCITRAL) instruments, and reference to three empirical studies on greenwashing, data portability, and Buy Now, Pay Later (BNPL). The trust-by-design model is the author's normative synthesis of these materials; it is proposed as a managerial framework and has not itself been empirically tested.

Key Recommendations

1. Adopt trust-by-design for consequential digital actions and assign a named business owner to every material AI application.
2. Preserve provenance and authentication records for high-consequence digital content and maintain substantiation files for material environmental claims.
3. Test whether data can be recovered from critical platforms and vendors, and require automated contracts to include authority, override, error, and dispute mechanisms.
4. Use decision-relevant friction in BNPL and embedded credit, meaning targeted disclosures, confirmations, or pauses at the point where a customer needs information to understand a credit commitment. Embedded credit refers to credit offered within a non-credit purchasing or payment journey.
5. Measure implementation through five operational trust indicators with defined numerators, denominators, data sources, and responsible units.

Summary of Facts

Philippine law already supplies important foundations for verifiable trust. The Electronic Commerce Act of 2000, gives legal recognition

to electronic documents subject to integrity, reliability, and authentication requirements (sec. 7), recognizes electronic signatures subject to statutory conditions (sec. 8), addresses authentication of electronic data messages and documents (sec. 11), and recognizes the formation and validity of electronic contracts (sec. 16). The Rules on Electronic Evidence, A.M. No. 01-7-01-SC, place the burden of proving authenticity on the proponent of an electronic document and specify methods of authentication (Rule 5, secs. 1–2); Rule 6 separately governs electronic signatures. *MCC Industrial Sales Corp. v. Ssangyong Corp.* (2007) is an important Supreme Court treatment of electronic evidence and the scope of the Electronic Commerce Act.

Comparative instruments illustrate parallel governance pressures. The EU Regulation (EU) 2024/1689 (Artificial Intelligence Act) addresses transparency for specified AI-generated or manipulated content. The EU Digital Markets Act (Regulation (EU) 2022/1925) imposes data-access and portability obligations on designated “gatekeepers,” a term of art for large digital platforms formally designated under that regulation; the term has no direct Philippine-law equivalent. The U.S. Federal Trade Commission Green Guides caution against broad, unqualified environmental-benefit claims. The 2024 Model Law (United Nations Commission on International Trade Law, 2024) supplies supplementary rules for automated contracting. These are comparative benchmarks, not statements that the same foreign duties automatically apply to Philippine firms.

The empirical evidence is narrower than the proposed framework and is used only to illuminate particular risks. In a preregistered experiment with a nationally representative Irish sample of 2,000 adults, Timmons et al. (2024) found that an anti-greenwashing intervention increased perceived greenwashing for two of three misleading advertisements, reduced brand trust and purchase

intentions, and also increased suspicion toward two of three genuine advertisements. Jeon et al. (2023) modeled data portability as reducing switching costs and altering firms' compatibility choices, while showing that consumer effects depend on market conditions. Using BNPL data from six large providers matched with de-identified credit records, Shupe and DeLuca (2025) examined persistent use, loan stacking, and consumers' aggregate unsecured debt exposure.

Statement of Issues

The Policy Brief addresses four linked questions. First, accountability: who owns the consequence of a material AI or automated decision? Second, authenticity: can the firm establish the source of consequential digital content and what happened to it? Third, substantiation: can the firm prove a material claim before communicating it? Fourth, contestability: can an affected person challenge, reverse, exit, or obtain redress from a consequential digital outcome? These four tests determine when the trust-by-design controls described below should apply.

Outline of Arguments

The legal subjects considered here are doctrinally distinct but share one managerial problem: digital capability is expanding faster than the firm's capacity to demonstrate the integrity of its digital actions. A compliance-led model responds to each legal requirement separately and offers familiar legal ownership, but it tends to be reactive and fragmented. A technology-led model emphasizes logging, watermarking, authentication, and monitoring; it scales well, but technical assurance alone does not, by itself, establish managerial responsibility or legal defensibility. Trust-by-design integrates legal, risk, technology, data, marketing, and product controls

around the four tests before consequential actions occur.

Control intensity should be proportionate, but it need not be expressed as a mathematical formula. Governance should become stronger as the magnitude, permanence, and reach of the potential consequence increase. A low-stakes internal AI summary may require little more than ordinary review. An automated credit denial affecting thousands of customers is different because the consequence is financially significant, difficult to reverse at scale, and widely distributed. This principle, referred to below as defensible proportionality, is that management should be able to explain why the controls chosen were commensurate with the consequences.

Discussion

Accountability requires an identifiable owner for every material AI application. AI can disperse responsibility among the business unit requesting a system, the technology team deploying it, the vendor supplying it, employees relying on its output, and executives authorizing its use. An AI material register should record the purpose, owner, provider, material data sources, level of autonomy, affected stakeholders, human-review requirement, escalation route, and conditions for suspension. Human oversight must be substantive: a reviewer needs the information, competence, time, and authority to reject or escalate an automated recommendation. Automation may distribute work; it should not distribute responsibility until responsibility disappears.

Authenticity requires firms to distinguish watermarking, provenance, and authentication. A watermark can signal artificial generation or manipulation; provenance records the source and history; authentication establishes whether an artifact is what its proponent claims it is. C2PA Content

Credentials provide mechanisms for recording information about digital assets and their derivation, rather than proving substantive truth (Coalition for Content Provenance and Authenticity, 2025). For consequential digital content, firms should preserve a defensible chain of custody from source and creation through AI involvement, modification, approval, publication, and preservation. Depending on risk, controls may include authenticated identities, source files, access logs, timestamps, version histories, digital signatures, cryptographic hashes, content credentials, and secure repositories.

Substantiation addresses the widening gap between how quickly claims are generated and how much work it takes to prove them. Environmental marketing illustrates the risk. Terms such as “green,” “sustainable,” and “carbon neutral” may conceal assumptions about product boundaries, supply chains, measurement periods, lifecycle effects, and offsets. The FTC Green Guides caution against broad, unqualified environmental-benefit claims and emphasize appropriate substantiation and qualification (Federal Trade Commission, 2012). Firms should require a substantiation file identifying the exact claim, supporting evidence, methodology, assumptions, period covered, responsible unit, approval, and reassessment date. The rule should apply whether the language originated from an employee, an agency, a consultant, or an AI system: no material claim without evidence to support it.

Contestability is the counterweight to frictionless automation. Firms may depend on digital platforms for customer access, transactions, analytics, payments, software, and data. Rather than assuming that nominal data ownership guarantees practical exit, a firm should conduct a data-exit test: if a critical provider became commercially, legally, technologically, or strategically unacceptable tomorrow, what information could be retrieved, in what format, within what period, at what cost, and with what loss of functionality? This translates the

the portability problem into an enterprise-resilience question without assuming that portability necessarily improves competition.

Smart contracts raise contestability differently. Electronic Commerce Act of 2000, sec. 16, recognizes electronic formation and validity, but technical execution does not by itself resolve authority, consent, mistake, interpretation, illegality, or remedy. Material automated transactions should therefore connect executable logic to an authoritative legal agreement, an authoritative record, an override mechanism, and a dispute mechanism. For BNPL and other embedded-credit journeys, decision-relevant friction should preserve the points at which a customer needs to understand the amount financed, payment schedule, material charges, missed-payment consequences, dispute resolution, and assistance before committing.

Implementation can be phased over 180 days. During days 0–30, firms should inventory material AI applications and consequential automated decisions. During days 30–60, they should assign business owners and risk classifications. During days 30–90, they should establish provenance and retention standards and require substantiation files for material environmental claims. During days 60–120, they should test data exit from critical providers, review automated contracts for errors, override, and dispute mechanisms, and review BNPL or embedded-credit journeys for informed choice and redress. During days 120–180, senior management should conduct a trust-failure simulation, meaning a tabletop exercise in which a plausible failure, such as a fabricated executive instruction or erroneous automated decision, is used to test authentication, escalation, suspension, evidence preservation, communication, and remediation.

Trust-by-design can itself become excessive. Too many approvals may discourage experimentation; indiscriminate retention can expose privacy and

cybersecurity; portability can create new security risks; provenance mechanisms may fail; and nominal human review can become automatic deference. Defensible proportionality addresses this risk by concentrating stronger controls on actions with greater magnitude, permanence, and reach and by requiring management to explain why the selected controls were appropriate.

Table 1. Operational Trust Indicators

Indicator	Definition	Data source/responsible unit
AI Accountability Coverage	Material AI apps with named owner, risk class, and review/escalation ÷ total material AI apps x 100	AI/risk registers Business + Risk/Legal
Provenance Coverage	High-consequence communication with reconstructable source, changes, approval, and final version ÷ total designated communications x 100	Content repo; approval logs IT/Records + Comms
Claims Substantiation Rate	Material environmental claims with current approved substantiation file ÷ total material environmental claims published x 100	Claims register; substantiation files Marketing + Sustainability/Legal
Data Exit Readiness	Critical providers passing export/recovery test within target time ÷ total	

	critical providers tested x 100	Vendor register; exit-test reports IT + Procurement
Digital Consumer Harm Rate	Confirmed complaints, reversals, remediation, or harmful outcomes attributable to relevant processes ÷ relevant transactions x 10,000	Complaint + transaction logs Products + Compliance

Conclusions

AI regulation, deepfake authentication, green claims, platform dependence, data portability, smart contracts, and BNPL appear to pose distinct compliance challenges. At the level of firm governance, however, they reflect a common institutional challenge: digital capability is expanding faster than the mechanisms for establishing the integrity of digital action. The response should not be seven independent compliance programs. It should be a common organizational capability built around four questions: Who owns the consequence? Can origin and material history be established? Can evidence support important claims? Can an affected person challenge, exit, reverse, or obtain redress?

A trustworthy firm should be able to reconstruct a consequential digital action from beginning to end: what happened, what information was used, whether AI participated, who possessed authority, what evidence supported the action, what controls applied, whether intervention remained possible, and what the firm did when the result was wrong. Trust has traditionally been treated as an intangible reputational asset. In the age of AI, that is no longer sufficient. Trust is becoming an auditable operating capability.

The firms best positioned to preserve it will build the capacity to prove it before they are required to do so.

References

Coalition for Content Provenance and Authenticity. (2025). *C2PA technical specification* (Version 2.2, May 2025). Publisher.

Department of Trade and Industry. 2024. “DTI Drives AI Innovation with National Roadmap 2.0 and CAIR Launch.”

European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/1925 on Contestable and Fair Markets in the Digital Sector (Digital Markets Act). Official Journal of the European Union L 265: 1–66.

European Parliament and Council of the European Union. 2024. Directive (EU) 2024/825 on Empowering Consumers for the Green Transition.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).

Federal Trade Commission. (2012). Guides for the use of environmental marketing claims (16 C.F.R. Part 260). Publisher.

Jeon, D.-S., Menicucci, D., & Nasr, N. (2023). Compatibility choices, switching costs, and data portability. *American Economic Journal: Microeconomics*, 15(1), 30–73. <https://doi.org/10.1257/mic.20200309>

MCC Industrial Sales Corp. v. Ssangyong Corp., G.R. No. 170633, 562 Phil. 390 (October 17, 2007) (Phil).

Electronic Commerce Act of 2000, Republic Act No. 8792 (2000) (Phil).

Shupe, C., & DeLuca, J. (2025). *Consumer use of buy now, pay later and other unsecured debt*. Consumer Financial Protection Bureau.

Supreme Court of the Philippines. (2001). *Rules on electronic evidence* (A.M. No. 01-7-01-SC). Supreme Court.

Timmons, S., Whelan, A., & Kelly, C. (2024). An experimental test of a greenwashing inoculation intervention in Ireland: Effects of ‘pre-bunking’ on identification, consumer trust and purchase intentions. *Sustainable Production and Consumption*, 47, 318–328. <https://doi.org/10.1016/j.spc.2024.03.030>

United Nations Commission on International Trade Law. (2024). *UNCITRAL Model Law on Automated Contracting*. United Nations.