California Management Review

ARTIFICIAL INTELLIGENCE

Escaping Technological Stockholm Syndrome: The Case for Artificial Integrity in AI Design

by Hamilton Mann

Image Credit | spainter_vfx

Helping Managers to Identify Risks That May Lead to Technological Stockholm Syndrome

INSIGHT | FRONTIER 10 Sep 2025

The adoption of digital technology cannot be reduced to a mere rational decision or a functional evolution of user practices. It represents a profound reconfiguration[1] of the individual's cognitive, social, and behavioral frameworks, driven by algorithmic and

prescriptive logics that override their own referential structures.

RELATED CMR ARTICLES

Vegard Kolbjørnsrud, "**Designing the Intelligent Organization: Six Principles For Human-AI Collaboration**," California Management Review, 66/2 (2023): 44–64.

This process of technological transition, far from being neutral, resembles a form of symbolic capture in which the individual, confronted with the violence of change, activates psychological defense mechanisms in response to what is perceived as an assault on their autonomy, free will, and identity integrity[2].

When adoption is deemed successful, it means that the initial defense structures have collapsed: the user has not only internalized the rules imposed by the technology but has also developed a form of emotional identification with it, reinterpreting the origin of the constraint as a chosen relationship[3, 4, 5].

At this stage, a new normative regime takes hold. This shift signifies the substitution of the individual's former frame of reference by that of the machine, now perceived as familiar and reassuring. The initial aggression is repressed, and the newly formed cognitive automatisms become objects of defense.

This phenomenon, which may be likened to Stockholm Syndrome in the relationship between humans and machines, involves a dislocation of cognitive referents followed by an emotional reconfiguration in which the victim ultimately comes to defend their technological aggressor[6].

The resulting cognitive subjugation is not a side effect, but rather a survival mechanism. It is fueled by the brain's attempts to reduce the stress caused by the intrusion of a foreign cognitive framework. This emotional rewriting enables the individual to maintain a form of internal coherence in the face of technological alienation.

The user's attention then shifts away from the initial violence to focus on the positive signals emitted by the machine: social validation, algorithmic gratification, and playful rewards. These stimuli activate the emotional confirmation bias and transform coercion into perceived benevolence.

Through a process of neural plasticity, the brain's circuits reorganize the perception of the relationship with the machine: what was once stress becomes normality; what was once domination becomes support; and what was once an aggressor becomes a companion. An inversion of the power dynamic occurs through the reconfiguration of the nucleus accumbens and the prefrontal cortex, anchoring a new constrained affective relationship[7].

This phenomenon represents one of the fundamental perils that artificial intelligence poses to humanity: the normalization of mental dependence as a vector of social acceptability. This is why it is not enough to design artificially intelligent systems; they must also be endowed with artificial integrity, as a safeguard for human cognitive sovereignty[8, 9, 10].

Some argue that digital technology contributes to the empowerment of individuals in vulnerable situations. However, this argument conceals a more troubling reality: technological dependence is often portrayed as regained autonomy, when in fact it is based on the prior collapse of identity-based self-defense mechanisms.

When these defenses are weakened or absent, adherence to technology is no longer a matter of choice but of necessity, eliminating the critical dimension of appropriation. In such cases, Technological Stockholm Syndrome does not emerge through reversal, but through a lack of resistance.

Even when technology aims to restore a relative sense of autonomy, the process of cognitive imposition remains active, facilitated by the weakness of the user's defense mechanisms. In a state of diminished resistance, the user adheres all the more quickly and deeply to the framework imposed by the machine.

In all cases, technology shapes a new cognitive environment. The distinction lies solely in the degree of integrity of the preexisting mental framework: the more robust this framework is, the stronger the resistance; the more it is degraded, the faster technological

infiltration occurs[4].

The paradox that prevents the systemic recognition of this syndrome is that of innovation itself. Perceived as inherently positive, it conceals its ambivalent potential: it can both emancipate and alienate, depending on the conditions under which it is adopted.

For artificial intelligence to strengthen our humanity without diluting it, it must be built not only on artificial cognitive capabilities but also on an ethics of integrity: a technology that respects the mental, emotional, and identity-based freedoms of individuals[8].

Thus, while some rush toward ever more artificial intelligence, it becomes imperative to advance toward a more essential form of intelligence: that of artificial integrity, the only true guarantor of our psychic sovereignty[9].

Technology can alleviate pain, reduce risk, and improve human existence. Yet no advancement should come at the cost of a cognitive debt that would undermine our ability to think for ourselves, and with it, our relationship to our own humanity.

The evaluation of artificial integrity in digital systems, particularly those incorporating artificial intelligence, must become a central requirement in any digital transformation. This entails identifying functional gaps, implementing corrective measures, and defining cognitive counter-mechanisms that preserve the human being in all their complexity[8, 9].

1. **Functional Diversion**: The use of technology for purposes or in roles not anticipated by its designer or the implementing organization can render both the intended logic of software use and internal governance mechanisms ineffective, thereby generating functional and relational confusion[11].
   Example: A chatbot initially designed to answer questions about company HR policies is repurposed as a substitute for human hierarchy in managing conflicts or assigning tasks.
2. **Functional** Loophole: The absence of necessary steps or features, due to their omission during development and thus their exclusion from the operational logic of the technology, creates a “functional loophole” in relation to the user’s intended use[12, 13].
   Example: A content generation technology (such as generative AI) that does not allow

direct export of the content into usable formats (e.g., Word, PDF, etc.) at the expected level of quality, thereby limiting or obstructing its operational use.

3. **Functional Safety**: The absence of safeguards, human validation steps, or informational messages during a system's execution of an action with irreversible consequences may result in outcomes that do not align with the user's intent[14].
   Example: A marketing technology automatically sends emails to a contact list without any mechanism to block the dispatch, request user confirmation, or trigger an alert in cases where a key criterion, such as verifying the correctness of the recipient list, has not been confirmed, thus compromising the safety and quality of the operation.
4. **Functional Alienation**: The creation of automatic behaviors or conditioned reflexes, akin to Pavlovian responses, can diminish or even eliminate the user's capacity for reflection and judgment, leading to a gradual erosion of their decision-making sovereignty[15].
   Example: Systematic acceptance of cookies or blind confirmation of system alerts by cognitively fatigued users.
5. **Functional Ideology**: Affective dependence on technology can lead to the weakening or neutralization of critical thinking, fostering the mental construction of an ideology that supports narratives of relativization, rationalization, or collective denial regarding the technology's performance or malfunction[16].
   Example: Justifying technological flaws or errors with arguments such as "It's not the tool's fault" or "The tool can't guess what the user forgot".
6. **Functional Cultural Coherence**: The antinomy and contradictory injunction between the logical framework imposed or influenced by technology and the behavioral values or principles promoted by organizational culture can generate internal tensions[17, 18].
   Example: A technological workflow that leads to the creation of validation and oversight teams reviewing the work of others within an organization that otherwise promotes and values team empowerment.
7. **Functional Transparency**: The absence or inaccessibility of transparency and explainability in decision-making mechanisms or algorithmic logics regarding how a technology operates can prevent the user from anticipating, overriding, or transcending the system's intent[19].
   Example: Candidate preselection performed by a technology that manages trade-offs and conflicts between user-defined criteria (e.g., experience, degrees, soft skills),

without making the weighting or exclusion rules explicitly visible, editable, or verifiable by the user.

8. **Functional Addiction**: The presence of features based on gamification, instant gratification, or micro-reward systems specifically designed to hack the user's motivational circuits can activate neurological reward mechanisms, stimulating repetitive, compulsive, and addictive behaviors that lead to emotional decompensation and self-reinforcing cycles[20].
   Example: Notifications, likes, infinite scroll algorithms, visual or auditory bonuses, and progression thresholds based on points, badges, levels, or scores used to exponentially and durably maintain user engagement.
9. **Functional Ownership**: The appropriation, reuse, or processing of personal or intellectual data by a technology—regardless of its public accessibility—without the informed, explicit, and meaningful consent of its owner or creator raises critical ethical and legal concerns[16].
   This includes, but is not limited to: personal data, creative works (texts, images, voice, videos, etc.), behavioral data (clicks, preferences, location, etc.), and knowledge artifacts (academic content, journalism, open-source material, etc.).
   Example: An AI model trained on images, texts, or voices of individuals found online, thereby monetizing someone's identity, knowledge, or creative work without prior authorization, and without any mechanism for explicit consent, licensing, or transparent attribution.
10. **Functional Bias**: A technology's inability to detect, mitigate, or prevent biases or discriminatory patterns, whether in its design, training data, decision-making logic, or deployment context, can result in unfair treatment, exclusion, or systemic distortion toward individuals or groups[9].
    Example: A facial recognition system whose performance is significantly less accurate for individuals with darker skin tones due to imbalanced training data, and which lacks functional safeguards against bias or any accountability mechanisms.

Given their interdependence with human systems, the ten functional gaps related to artificial integrity must be examined through a systemic approach, encompassing the nano (biological, neurological), micro (individual, behavioral), macro (organizational, institutional), and meta (cultural, ideological) levels.

The cost associated with the absence of artificial integrity in systems, whether or not they incorporate artificial intelligence, affects multiple forms of capital: human (skills, engagement, mental health), cultural (values, internal coherence), decisional (sovereignty, responsibility), reputational (stakeholder trust), technological (actual value of technologies), and financial (inefficiency, underperformance of investments, maintenance overruns, corrective expenses, litigation, lost opportunities, and value destruction).

This cost manifests as sustained value destruction, driven by unsustainable risks and an uncontrolled increase in the cost of capital required to generate returns (ROIC), ultimately turning technological investments into structural liabilities for a company's profitability and, consequently, its long-term viability.

A company does not adopt responsible digital transformation solely to meet societal expectations, but because its long-term performance depends on it, and because it thereby helps strengthen the living social fabric that sustains it and upon which it relies to grow.

References

1. Douglas H. Schultz and Michael W. Cole, "**Higher Intelligence Is Associated with Less Task-Related Brain Network Reconfiguration**," Journal of Neuroscience, 36/33 (August 2016): 8551–8561.
2. Richard P. Bagozzi, Fred D. Davis, and Paul R. Warshaw, "**Development and Test of a Theory of Technological Learning and Usage**," Human Relations, 45/7 (July 1992): 659–686.
3. Fred D. Davis, "**Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology**," MIS Quarterly, 13/3 (September 1989): 319–340.
4. Hartmut Rosa, Social Acceleration: A New Theory of Modernity (New York, NY: Columbia University Press, 2012).
5. Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis, "U**ser Acceptance of Information Technology: Toward a Unified View**," MIS Quarterly, 27/3 (September 2003): 425–478. .
6. Bessel van der Kolk, The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma (New York, NY: Viking, 2014).

7. Norman Doidge, The Brain That Changes Itself (New York, NY: Penguin Books, 2007).
8. Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (Oxford, UK: Oxford University Press, 2014).
9. Hamilton Mann, Artificial Integrity: The Paths to Leading AI toward a Human-Centered Future (Hoboken, NJ: Wiley, 2024).
10. Viswanath Venkatesh and Hillol Bala, "**Technology Acceptance Model 3 and a Research Agenda on Interventions**," Decision Sciences, 39/2 (May 2008): 273–315.
11. John Rooksby and Ian Sommerville, "**The Management and Use of Social Network Sites in a Government Department**," *Computer Supported Cooperative Work*, 21/4–5 (August 2012): 397–415.
12. Donald A. Norman, The Design of Everyday Things, revised and expanded edition (New York, NY: Basic Books, 2013).
13. Peter-Paul Verbeek, What Things Do: Philosophical Reflections on Technology, Agency, and Design (University Park, PA: Penn State Press, 2006).
14. Charles Perrow, Normal Accidents: Living with High-Risk Technologies (Princeton, NJ: Princeton University Press, 1984).
15. Nicholas G. Carr, The Shallows: What the Internet Is Doing to Our Brains (New York, NY: W. W. Norton & Company, 2010).
16. Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Pierre Chazerand, Virginia Dignum, et al., "**AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations**," Minds and Machines, 28/4 (December 2018): 689–707.
17. Arman Ghafoori, Manjul Gupta, Mohammad I. Merhi, Samrat Gupta, and Adam P. Shore, "**Toward the Role of Organizational Culture in Data-Driven Digital Transformation**," International Journal of Production Economics, 271 (May 2024): Article 109205.
18. Simon Alexander Wiese, Johannes Lehmann, and Michael Beckmann, "**Organizational Culture and the Usage of Industry 4.0 Technologies: Evidence from Swiss Businesses**," *arXiv* (December 2024).
19. Open Global Rights, "**Why Does Algorithmic Transparency Matter and What Can We Do About It?**" *Open Global Rights*, April 9, 2025.
20. Esmaeel Taghipour, Fatemeh Vizeshfar, and Nahid Zarifsanaiey, "**The Effect of Gamification-Based Training on the Knowledge, Attitudes, and Academic**

Achievement of Male Adolescents in Preventing Substance and Internet Addiction," *BMC Medical Education*, 23 (November 13, 2023): Article 860.

Hamilton Mann Follow

Hamilton Mann is Group VP of Digital at Thales and lecturer at INSEAD and HEC Paris. He is a globally recognized expert in AI for Good and was inducted into the Thinkers50 Radar as one of the Top 30 most prominent rising business thinkers. Mann is the author of "Artificial Integrity" (Wiley).