When Engagement Stops Pointing:
Goodhart's Law and the Collapse of Social Discovery

Flyxion

January 12, 2026

Abstract

Goodhart's law states that when a measure becomes a target, it ceases to be a good measure. This principle offers a unifying explanation for a range of failures observed in engagement-driven social platforms. Metrics such as likes, follows, comments, and tags were originally introduced as proxies for interest, relevance, and social connection. Once these proxies became targets for optimization and monetization, their informational value degraded.

This essay argues that Facebook's engagement-optimized architecture has rendered the platform increasingly ineffective for discovery-based tasks such as dating, collaboration, and interest alignment. Practices including follow-for-follow networks, engagement farming, broadcast tagging, and identity performance emerge not as abuses but as rational adaptations to metric pressure. As a result, signals that once pointed toward shared interest or interpersonal address now primarily indicate algorithmic success.

Drawing analogies from classroom assessment, early online forums, and contemporary platform dynamics, the paper shows how metric capture produces defensive behavior, false addressability, and the erosion of trust. The failure is structural rather than moral: users adapt sensibly to incentives that reward reaction over relevance. The essay concludes that restoring meaningful discovery requires limiting what can be optimized and reestablishing a separation between measurement, reward, and social address.

1 Introduction

Goodhart's law states that when a measure becomes a target, it ceases to be a good measure. Originally formulated in the context of economic policy, the principle has proven broadly applicable to complex adaptive systems in which quantitative proxies are used to guide behavior. Social media platforms represent one such system. Metrics such as likes, follows, comments, and shares were initially introduced as rough indicators of interest, relevance, and social connection. Over time, however, these indicators have been elevated from descriptive signals to optimization targets.

Facebook, in particular, has built an ecosystem in which engagement functions simultaneously as a measure of success, a mechanism of distribution, and a unit of monetization. Content is promoted not on the basis of shared interest or interpersonal relevance, but according to its ability to generate measurable reactions. As creators adapt to these incentives, the meaning of engagement itself changes.

A like no longer reliably signals interest; a follow no longer implies affinity; a comment no longer implies conversation.

This essay argues that Facebook's engagement-driven architecture has rendered the platform increasingly ineffective for one of its original purposes: helping people find and connect with others who share their interests. The proliferation of follow-for-follow networks, engagement groups, and broadcast tagging practices such as `@followers` has produced a dense field of social signals that appear meaningful while conveying little actionable information. What remains is a system saturated with attention but impoverished in address, where users are constantly notified that they are being "engaged" without being genuinely spoken to.

By examining Facebook through the lens of Goodhart's law, this paper seeks to show that the platform's current failures are not primarily the result of malicious actors or individual bad behavior, but of structural incentive alignment. When engagement becomes the target, interest disappears as a signal.

2 Why Dating and Serious Collaboration Fail Under Engagement Optimization

Systems designed for dating, partnership, or collaboration depend on a specific kind of signal fidelity. At minimum, they require that identity be costly to fake, that expressions of interest be directionally meaningful, and that attention be scarce enough to indicate genuine intent. Facebook's engagement-optimized architecture undermines all three conditions.

As engagement became the dominant performance metric, identity itself was transformed into a vehicle for attention capture. The platform rewards content and accounts that elicit rapid, affective responses—especially sexual curiosity, novelty, or outrage—regardless of whether the underlying identity is authentic. This incentive structure predictably selects for fake or misleading profiles, recycled images, and exaggerated personas designed to trigger engagement rather than represent a real individual with stable interests or goals.

The result is a proliferation of accounts whose purpose is not connection but extraction: eliciting sexual responses, harvesting comments, driving clicks, or inflating visibility metrics. These accounts need not succeed in deception for long; they only need to perform well momentarily under the engagement model. From the perspective of Goodhart's law, identity has ceased to function as a reliable measure and has become a target to be optimized.

This failure mode is especially visible in attempts to use Facebook for dating. Dating requires mutual interest grounded in relatively truthful self-presentation and selective address. Instead, users encounter an environment saturated with clickbait sexuality and performative flirtation, where responses signal algorithmic success rather than interpersonal intent. A sexual reaction no longer indicates attraction to a person; it indicates susceptibility to a stimulus. As a consequence, genuine signals of interest are drowned out by optimized noise.

The same dynamics undermine the platform's usefulness for forming business ventures or long-term collaborations. Productive collaboration depends on discovering people with overlapping goals, complementary skills, and compatible working styles. Engagement metrics, however, are indifferent to these qualities. They privilege visibility over relevance and reward rhetorical performance over substantive alignment. Groups form not around shared objectives, but around reciprocal engagement strategies that inflate apparent activity while conveying little information about competence or intent.

In this context, attempts to use Facebook as a discovery tool for dating or serious projects fail not because users behave dishonestly, but because honesty is systematically disfavored. Truthful self-representation is slower, less reactive, and less immediately engaging than optimized performance. Under sustained selection pressure, the platform evolves toward identities that maximize response rather than correspondence with reality.

Goodhart's law predicts this outcome precisely. When engagement becomes the target, signals that once pointed toward shared interest, mutual attraction, or common purpose are repurposed to maximize measurable reaction. What remains is a dense field of attention without trust, visibility without address, and interaction without orientation toward real-world commitment.

3 Why Moderation Cannot Repair Metric Capture

It is tempting to interpret the failures of engagement-driven platforms as problems of enforcement rather than structure. From this perspective, fake accounts, sexual clickbait, and deceptive identity practices appear as pathologies that could be corrected through better moderation, stricter rules, or more advanced detection systems. This view misunderstands the nature of the failure. When Goodhart's law applies, moderation cannot restore the original meaning of a signal, because the signal has already been repurposed by the incentive structure itself.

Moderation operates by enforcing boundaries around unacceptable behavior. Engagement optimization, by contrast, reshapes what counts as successful behavior within those boundaries. As long as visibility and distribution are governed by engagement metrics, actors will continue to search for forms of expression that maximize reaction while remaining technically compliant. Sexualized imagery that stops just short of explicitness, misleading profile photos that avoid provable impersonation, and vague or generic self-descriptions all thrive precisely because they exploit the gap between formal rules and algorithmic reward.

Crucially, moderation evaluates content at the level of rule violation, not at the level of signal fidelity. A fake account need not violate any explicit policy to degrade the informational value of identity on the platform. A post designed to elicit sexual curiosity need not be pornographic to function as clickbait. A network of reciprocal engagement accounts need not be coordinated explicitly to distort discovery. These behaviors are not aberrations; they are stable equilibria under the current optimization regime.

As moderation efforts intensify, they often increase pressure toward more subtle forms of optimization rather than eliminating them. This produces an arms race in which surface-level abuses are pruned while deeper distortions persist. The platform may remove the most obvious scams, but it cannot reintroduce trust into identity signals without changing what the system rewards. In effect, moderation treats symptoms while the disease continues to operate unimpeded.

The core problem is that engagement metrics are being asked to perform incompatible roles. They are simultaneously used as indicators of relevance, mechanisms of distribution, and inputs to monetization. Once these functions are entangled, no amount of content policing can recover the original semantics of social interaction. The measure has become the target, and moderation cannot reverse that transformation.

From the perspective of Goodhart's law, the failure of moderation is not accidental but inevitable. Enforcement can constrain behavior at the margins, but it cannot restore meaning to a signal whose informational role has been structurally displaced. To do that would require decoupling engagement from success itself—an intervention far deeper than any moderation policy can reach.

4 Why Older Forums and Classifieds Enabled Real Discovery

Before the dominance of engagement-optimized social platforms, online discovery often occurred in spaces that were technically simpler but semantically richer. Bulletin boards, mailing lists, forums, and early classifieds lacked algorithmic feeds, monetized attention loops, and real-time performance metrics. Yet they were frequently more effective at enabling dating, collaboration, and interest-based connection. This was not despite their limitations, but because of them.

In these earlier systems, visibility was not dynamically allocated according to engagement. Posts were typically displayed in chronological or categorical order, and attention was governed by deliberate navigation rather than algorithmic amplification. As a result, expressive success depended less on provoking immediate reaction and more on accurately signaling relevance to a specific audience. A message persisted not because it performed well, but because it was placed where interested readers could intentionally find it.

Identity in these environments was also more tightly coupled to cost. Creating and maintaining a presence required sustained participation, recognizable authorship, and reputational continuity within a bounded community. While anonymity was common, it was contextual rather than opportunistic. Pseudonyms accrued meaning through repeated interaction, and deceptive identities were harder to sustain because there was no global engagement metric to exploit. Trust emerged slowly, through pattern and history, rather than through visibility spikes.

Dating and collaboration benefited directly from this structure. Classified ads and interest-specific forums encouraged explicit self-description and goal articulation. Because there was little incentive to maximize reaction, exaggeration and sexual clickbait conferred few advantages. A misleading post might attract attention briefly, but it would not be algorithmically amplified, nor would it persist without follow-up. In this sense, the systems penalized noise naturally, not through enforcement, but through indifference.

Crucially, these spaces preserved the distinction between address and broadcast. A post in a forum was directed at a known readership defined by topic or purpose. Replies were situated responses rather than generic engagement tokens. This made expressions of interest legible: a reply meant someone had read and chosen to respond, not merely reacted. Discovery functioned because signals pointed reliably toward shared interest rather than toward attention capture.

The contrast with contemporary platforms is instructive. Modern social networks are more powerful technologically but poorer informationally. By collapsing navigation, discovery, and monetization into a single engagement-driven feed, they eliminate the environmental constraints that once preserved signal meaning. What older systems lacked in scale, they compensated for in selectivity. What they lacked in polish, they made up for in trust.

From the perspective of Goodhart's law, early online spaces avoided metric capture not through foresight, but through structural simplicity. Because attention was not a target, it remained informative. Because engagement was not monetized, it remained scarce. These conditions enabled genuine discovery, even in the absence of sophisticated algorithms. Their disappearance marks not technological progress, but a loss of semantic resolution.

5 Classrooms as Early Sites of Metric Capture

The dynamics described in this essay are not unique to digital platforms. They appear wherever performance is evaluated through simplified metrics that stand in for deeper goals. Classrooms provide an instructive early example. Educational systems frequently rely on tests, rubrics, or behavioral rules as proxies for learning, understanding, and intellectual development. When these proxies are elevated to targets, they begin to shape behavior in ways that undermine their original purpose.

One familiar manifestation of this process is teaching to the test. When standardized assessments become the primary measure of success, instruction adapts accordingly. Teachers may narrow curricula, emphasize test-taking strategies, or discourage exploratory reasoning that does not map cleanly onto evaluated outcomes. Students, in turn, learn to optimize for performance rather than comprehension. Mastery becomes indistinguishable from compliance with the metric.

A related but subtler effect occurs when classroom rules vary idiosyncratically across teachers without shared rationale or explanation. In such environments, students quickly learn that success depends less on general principles of reasoning or conduct than on local, person-specific constraints. Certain words, questions, or behaviors may be acceptable in one classroom and penalized in another. Over time, students internalize not a coherent model of good participation, but a fragmented set of avoidance strategies: do not say this here, do not ask that there, do not behave this way with this authority figure.

This adaptation is rational. Faced with inconsistent or opaque rules, students optimize for survival within each context. However, the result is a form of learned fragmentation. Instead of generalizing from principles, students learn to treat each classroom as a separate game with its own hidden scoring system. The measure of success becomes the avoidance of penalties rather than the pursuit of understanding.

By contrast, classrooms that involve students in rule formation or explicitly explain the function of constraints produce different outcomes. When rules are justified in terms of shared goals—facilitating discussion, ensuring fairness, enabling focus—students are more likely to internalize transferable norms. In these cases, evaluation criteria remain connected to their underlying purpose. The metric does not replace the goal; it points back to it.

This contrast illustrates the core mechanism of Goodhart's law at a human scale. When rules or assessments are treated as targets rather than instruments, participants adapt by optimizing locally and defensively. Learning becomes brittle and context-bound. When measures are transparently linked to their function, participants can generalize, internalize, and cooperate in maintaining the system itself.

The classroom example matters because it reveals that metric capture is not primarily a technological problem. It is a coordination problem. Social platforms replicate the worst version of the classroom dynamic at scale: opaque rules, inconsistent enforcement, and performance metrics disconnected from shared understanding. Users, like students, adapt by learning what to avoid rather than what to mean. The result is not collective learning, but widespread strategic silence, performative speech, and context-specific self-censorship.

6 AI Platforms, Engagement Arms Races, and the Collapse of Identity Sorting

The dynamics described in this essay are not confined to legacy social media platforms. They are increasingly visible in contemporary AI-driven systems, including generative media tools and conversational interfaces that incorporate social feeds, discovery surfaces, or public timelines. In these environments, the same pattern reappears: outputs are ranked, promoted, and surfaced according to engagement proxies rather than alignment with user intent, domain interest, or sustained work.

Recent examples include generative video, music, and text platforms that increasingly resemble short-form social media feeds. Instead of functioning primarily as tools for creation, study, or exploration, they present users with streams of visually or emotionally striking content optimized for rapid reaction. The result is a flood of clickbait-style imagery, novelty-driven demonstrations, and performative outputs that compete for attention in an engagement arms race reminiscent of TikTok-style dynamics. What is rewarded is not usefulness or relevance, but immediate impressiveness.

This shift is especially problematic because these platforms implicitly promise something different. Users approach them as tools for work, learning, or creative development. They expect to be able to explore mathematics, music, programming, psychology, or other domains in depth. Instead, they encounter a flattened surface where all outputs are rendered comparable by the same engagement criteria. A serious mathematical visualization, a novelty music clip, and a sensationalized video demonstration are placed in direct competition for attention, despite serving fundamentally different purposes.

This convergence produces a collapse of identity sorting. In order to function well, individuals require multiple, context-sensitive identities: one for technical study, one for artistic exploration, one for professional collaboration, one for local coordination, and so on. Each of these contexts employs different norms, vocabularies, and evaluative criteria. Mathematical reasoning tolerates abstraction and contradiction differently than psychology; musical discourse values expressiveness that would be inappropriate in formal logic; local event discovery depends on geographic specificity rather than global visibility.

Engagement-driven feeds erase these distinctions. When a single identity is forced to perform across incompatible domains, users adapt by diluting expression. Language becomes generic, outputs become broadly palatable, and interests are compressed into a lowest common denominator that travels well across the feed. As in the classroom case, participants learn what to avoid rather than how to articulate purpose. The system trains them to minimize friction rather than maximize meaning.

The problem extends beyond intellectual domains to practical coordination. Individuals who commute between cities, participate in multiple local communities, or balance remote and in-person work require fine-grained contextual discovery. Local events, collaborators, or study groups are valuable precisely because they are not globally engaging. An engagement-optimized system systematically suppresses this information, favoring content that travels widely over content that is locally useful.

From the perspective of Goodhart's law, these AI platforms are reproducing the same error under a new technological guise. Engagement is treated as a universal measure of value across contexts that are not commensurable. As a result, the systems optimize toward spectacle rather than support, toward performance rather than orientation. The tools remain powerful, but their interfaces increasingly resemble attention markets rather than workspaces.

The consequence is not merely distraction, but a deeper loss of navigability. When users cannot reliably separate their interests, roles, and locations, discovery fails even in the presence of abundant content. What is missing is not intelligence or capability, but structural respect for plurality: the recognition that different kinds of work, inquiry, and coordination require different evaluative regimes, and that no single engagement metric can serve them all without erasing their meaning.

7 A Shared Failure Mode Across Contemporary Platforms

The relevance of AI-driven feeds to the broader argument of this essay is not incidental. They demonstrate that the failures commonly attributed to legacy social media platforms are not historical anomalies or the result of particular corporate choices, but expressions of a general failure mode that emerges whenever engagement is treated as a universal proxy for value. The same structural logic appears across systems regardless of whether the content is user-generated, algorithmically curated, or produced by generative models.

In each case, discovery surfaces are optimized for reaction rather than orientation. Whether a feed presents short videos, generated images, music clips, or conversational outputs, the governing criterion remains what provokes response and retains attention. The system does not distinguish between contexts of use. It does not ask whether the user is studying mathematics, composing music, seeking collaborators, or looking for local events. It measures only performance.

This continuity matters because it reveals that the problem is not medium-specific. Generative AI systems inherit the same distortions once they incorporate ranking, public visibility, or social comparison. What begins as a tool for exploration or production acquires a performative layer, and outputs gradually shift toward novelty, spectacle, and generalized appeal. This occurs even when users are seeking precision, depth, or domain-specific work.

Seen in this light, platforms like Facebook are not exceptions but early and visible instances of metric capture. Their failures prefigure those now appearing in newer systems that promise intelligence, creativity, or productivity while quietly reorganizing discovery around engagement. Signals that once helped users locate people, ideas, or opportunities increasingly indicate only algorithmic success.

This shared pattern sets the stage for examining how such dynamics manifest in more familiar settings, from classrooms and assessment systems to dating, collaboration, and everyday social coordination. In each case, the substitution of performance metrics for contextual understanding produces the same outcome: increased activity paired with diminished meaning.

8 False Addressability and the Erosion of Social Meaning

One of the most psychologically disruptive consequences of engagement-optimized systems is the emergence of false addressability: situations in which users are repeatedly signaled as being addressed, acknowledged, or included, without actually being the intended recipient of communication. This phenomenon is structurally analogous to the classroom dynamics described above, but operates at scale and with greater cognitive intensity.

On platforms such as Facebook, notifications, mentions, and broadcast tags—most notably `@followers`—create the appearance of direct address. The user is alerted, visually and temporally, as though someone has chosen to speak to them. In reality, these signals are often indiscriminate broadcasts optimized for reach rather than communication. The user is not being addressed as a person, but as a member of a monetizable audience segment.

This distinction matters because address carries implicit commitments. To address someone is to recognize their presence, to orient speech toward them, and to accept a minimal responsibility for relevance. Broadcast tagging collapses this distinction. It allows a single utterance to masquerade as many individual conversations simultaneously, without incurring the cost of specificity. From the perspective of Goodhart's law, address has become a target: the appearance of being spoken to is optimized, while the substance of being spoken with disappears.

Users adapt predictably. Over time, they learn that notifications are unreliable indicators of relevance. A tag no longer means that someone has something to say to them; it means that someone is attempting to activate engagement. The rational response is desensitization. Messages are skimmed, ignored, or dismissed preemptively. Ironically, this makes genuine attempts at communication harder to detect, as they are filtered through the same degraded signaling channel.

This dynamic mirrors the classroom case in which students learn to avoid saying certain things in certain contexts without understanding why. Here, users learn to discount being "addressed" without being able to distinguish between sincere and performative address. The platform trains its participants not to listen, but to defend themselves against interruption. Attention becomes a scarce resource to be protected rather than a medium for connection.

The psychological cost of false addressability is cumulative. Repeated exposure to signals that imply recognition without delivering it produces a low-grade form of social noise. Users experience constant reminders of supposed relevance—likes, tags, mentions—that fail to translate into meaningful interaction. This creates a feedback loop in which people feel both overstimulated and unseen, engaged and ignored.

From a structural standpoint, false addressability represents the final stage of metric capture. Not only have interest and identity ceased to function as reliable signals, but address itself has been hollowed out. Communication no longer orients toward shared understanding or response; it orients toward measurable activation. In such an environment, silence becomes rational, specificity becomes costly, and genuine dialogue becomes increasingly rare.

What remains is a system rich in notifications and poor in conversation, saturated with signals that resemble social connection while systematically undermining its possibility.

9 From Local Optimization to Systemic Failure

Across classrooms, social platforms, and digital marketplaces, the same structural pattern repeats. A complex human goal—learning, connection, collaboration, trust—is approximated by a measurable proxy. That proxy is then elevated into a target. Participants adapt rationally. Over time, the behaviors that maximize the metric diverge from the behaviors the metric was originally meant to represent. The system does not merely fail; it succeeds at optimizing the wrong thing.

In classrooms, this process produces students who learn how to avoid penalties rather than how to reason. In dating environments, it produces performances that elicit reaction rather than attraction. In professional contexts, it produces visibility without alignment and networking without cooperation. On social platforms, it produces engagement without interest, address without communication, and identity without accountability.

What makes these failures difficult to correct is that they emerge from coordination rather than malice. No single participant needs to act deceptively for the system to degrade. Each actor responds to the incentives presented to them, and the aggregate outcome reflects those incentives with mechanical fidelity. Attempts to correct the system through moderation, etiquette, or exhortation fail because they leave the underlying optimization regime intact.

The cumulative effect is a shift from shared meaning to defensive strategy. Participants learn not how to contribute, but how to avoid missteps. Speech becomes generic to minimize risk. Identity becomes stylized to maximize reach. Silence becomes preferable to specificity. The system trains its users to treat every context as adversarial, every signal as unreliable, and every notification as suspect.

At this stage, the platform's original promise collapses under its own instrumentation. A tool designed to connect people becomes incapable of reliably indicating who is interested in whom or in what. Discovery degenerates into trend-following. Address becomes broadcast. Trust becomes untenable because the signals that once supported it no longer point toward anything stable.

Goodhart's law does not merely diagnose this outcome; it explains its inevitability. Once a proxy is converted into a target at scale, the system reorganizes around that target. The only durable remedy is not better policing or more sophisticated metrics, but structural restraint: limiting what can be optimized, preserving the cost of signaling, and maintaining a separation between measurement and reward.

Absent such restraint, platforms will continue to produce the same paradoxical result: ever-increasing engagement paired with ever-decreasing meaning.

10 How GitHub Preserves Identity While Allowing Plurality

Not all large-scale digital platforms collapse under engagement optimization. GitHub provides a useful counterexample, not because it is immune to social dynamics, but because its core architecture resists metric capture in key ways. In particular, it preserves stable identity while allowing users to express plurality through clearly separated contexts of work.

On GitHub, identity is fixed and relatively costly to change. Usernames are persistent, histories are public, and actions are recorded as part of a long-lived contribution graph. This persistence discourages opportunistic identity performance and supports accountability over time. At the same time, the platform does not force all activity into a single undifferentiated feed. Instead, it allows users to create and participate in multiple repositories, each with its own purpose, language, norms, and audience.

These repositories function as contextual identity partitions. A user may maintain a project written in Spanish using Python and heavy set-theoretic formalism, appealing to a mathematically inclined and linguistically specific group of collaborators. The same user may also contribute to a project written in Arabic on the history of Europe, centered on archival research and $\LaTeX$ composition. These activities coexist under a single identity without being collapsed into a single performance surface. Each repository is discoverable by those who care about its content, not by those responding to its immediate engagement potential.

Crucially, GitHub does not require these contexts to compete with one another for attention. There is no global engagement feed in which unrelated projects are ranked against each other by reaction metrics. Discovery occurs through intentional navigation: search, topic tags, dependency graphs, and direct links. Attention is pulled by interest rather than pushed by algorithmic amplification. As a result, signals remain legible. A star, a fork, or a pull request indicates domain-specific relevance, not generalized popularity.

This structure reduces the need for multiple accounts. In principle, a single GitHub identity can support professional work, academic research, hobby projects, and experimental exploration without confusion. The platform already provides the necessary separation through repositories, organizations, and contribution histories. When users nonetheless feel pressure to create separate accounts, the cause is rarely technical. More often, it reflects external institutional pressures.

Hiring agencies and large corporations tend to be image-protective and myopic. They often treat identity as a brand rather than a history, scanning profiles for perceived risk rather than contextual fit. Under these conditions, individuals may self-segregate, hiding exploratory or nonconforming work behind alternate accounts. This behavior is a rational response to institutional signaling failures, not an indictment of the platform's design.

From the perspective of this essay, GitHub succeeds where engagement-driven platforms fail because it respects the plurality of human activity without collapsing it into a single metric space. It allows one identity to speak in many languages, across many domains, without forcing those expressions to compete for the same kind of attention. Meaning is preserved not through moderation or moral exhortation, but through structural separation of contexts and restraint in what is optimized.

The contrast underscores a central claim of this paper: discovery works when systems preserve stable identity while allowing contextual differentiation. It fails when identities are flattened, signals are monetized, and all expression is forced to perform under a single evaluative regime.

11 Formal Expression, Generalizability, and the Fragility of Portfolios

One reason technical work has historically generalized across languages, cultures, and domains is that it often included formal components alongside prose. Algorithms, mathematical notation, type signatures, and explicit specifications constrain interpretation in ways that natural language alone cannot. While prose is indispensable for motivation and explanation, formal elements provide a shared reference frame that resists ambiguity. A proof, a formula, or a well-defined algorithm communicates intent more reliably than description, particularly across linguistic and cultural boundaries.

For much of the history of software development and technical collaboration, the difficulty of producing such artifacts served an unintended but useful function. Learning a programming language, setting up a development environment, understanding version control, and expressing ideas in code imposed real costs. These costs filtered participation toward individuals willing to invest time in understanding the systems they were using. As a result, repositories functioned not only as collections of working artifacts, but as credible signals of comprehension. A project's existence implied a minimum level of conceptual engagement.

This filtering effect is now weakening. The rise of natural-language-driven programming and generative code systems lowers the barrier to producing functional outputs without requiring corresponding understanding. While this has clear benefits for accessibility and rapid prototyping, it also alters the informational value of technical artifacts. A working project no longer reliably indicates that its author understands the language, the algorithms, or even the problem domain involved.

This shift has consequences for generalizability. When work is expressed primarily in natural language or generated code, its meaning becomes dependent on interpretation rather than structure. Ambiguities that formal notation would have resolved remain latent. Two collaborators may believe they are aligned while operating under incompatible assumptions, only discovering the mismatch downstream. The loss is not merely precision, but shared orientation.

It also has consequences for trust. As the cost of producing plausible technical projects approaches zero, portfolios lose their discriminative power. Repositories that once functioned as evidence of sustained competence become vulnerable to imitation and misuse. In the extreme case, working projects can be incorporated into phishing or social engineering campaigns, lending technical credibility to malicious activity. The artifact performs, but it does not testify.

From this perspective, platforms like GitHub are not immune to the dynamics described elsewhere in this essay. They remain structurally better than engagement-driven feeds, but they are not insulated from enshittification. If visible output replaces demonstrated understanding as the primary signal of value, the platform's role as a discovery and evaluation mechanism will degrade. The danger is not that automation exists, but that it collapses the distinction between producing an artifact and understanding it.

The implication is not that formalism should replace natural language, but that it must accompany it. Algorithms, equations, and explicit specifications act as anchors that stabilize meaning across contexts. They make work legible to collaborators who do not share the same linguistic or cultural background, and they preserve generalizability when prose alone would fragment.

If discovery systems are to remain useful for serious work, they must continue to reward forms of expression that resist easy imitation and preserve semantic commitment. Otherwise, even the best-designed platforms will drift toward the same failure mode: abundant output, diminished signal, and the gradual erosion of trust in what artifacts are supposed to represent.

12 Conclusion

This essay has argued that the widespread failures of contemporary digital platforms are not accidental, moral, or primarily technological, but structural. By applying Goodhart's law across domains as varied as social media, education, AI systems, and professional portfolios, the paper has shown how the elevation of proxies into targets systematically degrades the very capacities those proxies were meant to support. Engagement, visibility, performance, and output were introduced as imperfect measures of interest, relevance, learning, or competence. Once optimized and monetized, they ceased to point reliably toward those underlying goals.

In the case of Facebook and similar platforms, engagement-driven optimization hollowed out the mechanisms of discovery. Signals that once helped users find people with shared interests, goals, or local relevance were repurposed to maximize reaction. Dating devolved into clickbait sexuality, collaboration into performative networking, and address into broadcast. The proliferation of fake or exaggerated identities, follow-for-follow economies, and false addressability did not represent abuse of the system but its logical equilibrium. Users adapted rationally to incentives that rewarded reaction over orientation, producing an environment saturated with attention and starved of meaning.

The classroom analogy demonstrated that these dynamics predate digital platforms. Whenever evaluation criteria are opaque, inconsistent, or disconnected from shared purpose, participants learn defensive optimization rather than generalizable understanding. Students learn what to avoid rather than what to internalize. The same pattern reappears online at scale: users learn to minimize risk, flatten expression, and treat each context as adversarial. Speech becomes generic, silence becomes rational, and trust erodes not through malice but through adaptation.

The analysis of AI-driven feeds extended this argument to newer systems that promise intelligence, creativity, or productivity. Once these tools adopt engagement-based discovery surfaces, they inherit the same failure mode. Outputs optimized for spectacle crowd out work optimized for depth. Distinct domains of inquiry—mathematics, music, psychology, local coordination—are forced into a single evaluative regime despite requiring incompatible norms and languages. Identity sorting collapses, and with it the user's ability to navigate meaningfully across roles, interests, and locations.

GitHub was examined as a partial counterexample, illustrating how stable identity combined with contextual separation can preserve signal fidelity. By allowing multiple repositories under a single identity, the platform enables plurality without collapse. Discovery remains interest-driven rather than performance-driven, and artifacts retain evidentiary value because they are costly to produce and situated within clear contexts. Yet even this architecture is not immune. The rise of natural-language programming and generative code threatens to decouple visible output from understanding, weakening portfolios as signals of competence and opening the door to imitation, misrepresentation, and exploitation.

Across all these cases, the same structural lesson emerges. Systems fail when they conflate measurement with reward, proxy with purpose, and visibility with value. Moderation, enforcement, and improved metrics cannot resolve this failure because they operate downstream of incentive alignment. Once a proxy becomes a target, the system reorganizes around it with mechanical inevitability.

The implication is not that measurement, automation, or large-scale coordination are inherently corrosive. It is that meaningful discovery depends on restraint. Signals must remain costly enough to testify, contexts must remain differentiated enough to orient, and identities must remain stable enough to accrue history without being flattened into brands. Formal expression—algorithms, mathematics, explicit specifications—matters not as elitism, but as a way of anchoring meaning across linguistic, cultural, and interpretive boundaries.

Ultimately, this paper has argued that discovery is not a problem of scale or intelligence, but of semantics. When systems respect plurality, preserve context, and limit what can be optimized, signals continue to point. When they do not, engagement rises while meaning collapses. Goodhart's law does not merely warn against bad metrics; it reveals a fundamental constraint on the design of social and informational systems. Ignoring that constraint does not produce neutral failure. It produces systems that function smoothly, profitably, and at scale—while systematically undermining the purposes they were built to serve.

References

[1] C. A. E. Goodhart. Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics*, Reserve Bank of Australia, 1975.

[2] D. T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, 1979.

[3] M. Strathern. *"Improving Ratings": Audit in the British University System. European Review*, 5(3):305–321, 1997.

[4] J. Z. Muller. *The Tyranny of Metrics.* Princeton University Press, Princeton, 2018.

[5] C. Doctorow. The enshittification of TikTok. *Pluralistic*, January 2023.

[6] C. Doctorow. Social quitting. *Pluralistic*, May 2023.

[7] C. Doctorow. *The Internet Con: How to Seize the Means of Computation.* Verso Books, London, 2024.

[8] S. Zuboff. *The Age of Surveillance Capitalism.* PublicAffairs, New York, 2019.

[9] N. Srnicek. *Platform Capitalism.* Polity Press, Cambridge, 2017.

[10] Z. Tufekci. Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journal*, 13:203–218, 2015.

[11] Z. Tufekci. YouTube, the great radicalizer. *The New York Times*, March 10, 2018.

[12] Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press, New Haven, 2006.

[13] J. Lanier. *Ten Arguments for Deleting Your Social Media Accounts Right Now.* Henry Holt, New York, 2018.

[14] S. Bowles. Policies designed for self-interested citizens may undermine "the moral sentiments": Evidence from economic experiments. *Science*, 320(5883):1605–1609, 2008.

[15] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press, Cambridge, 1990.

[16] A. Kohn. *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes.* Houghton Mifflin, Boston, 1999.

[17] E. L. Deci and R. M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior.* Plenum Press, New York, 1985.

[18] G. A. Akerlof and R. J. Shiller. *Phishing for Phools: The Economics of Manipulation and Deception.* Princeton University Press, Princeton, 2015.

[19] A. O. Hirschman. *Exit, Voice, and Loyalty.* Harvard University Press, Cambridge, MA, 1970.

[20] J. C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, New Haven, 1998.