Viewpoint

Multilayered Epistemic Disruption in AI-Driven Health Misinformation: Conceptual Framework and Viewpoint

Muzaffer Malkoç, MA

Istanbul Medipol University, Beykoz, Istanbul, Türkiye

Corresponding Author:

Muzaffer Malkoç, MA
Istanbul Medipol University
Kavacık Güney Kampüsü, Göztepe Mah.
Beykoz, Istanbul 34810
Türkiye
Phone: 90 5334162689
Email: muzaffermalkoc@medipol.edu.tr

Abstract

Generative AI has transformed the health information ecosystem by enabling scalable, sophisticated health misinformation production at near-zero marginal cost. Current literature addresses AI's role in health misinformation predominantly through a binary threat detection framework, systematically overlooking the structural, multilayered mechanisms through which AI simultaneously embeds false claims across intersecting human trust systems. This paper introduces the Multilayered Epistemic Disruption Framework (MEDF), which conceptualizes how AI-driven health misinformation structurally undermines public trust through four interdependent layers of cognitive and institutional disruption: discursive (clinical language shielding: fluent medical terminology and fabricated citations deployed as credibility signals), biometric (embodied authority transfer: deepfake appropriation of real clinicians' faces and voices), temporal (the synthetic chorus effect: near-simultaneous fabrication of apparently independent corroborating sources), and systemic (structural epistemic erosion: cumulative macro-level collapse of trust in medical institutions). Adopting a socioecological and structural epistemic approach, this viewpoint synthesizes empirical findings from communication psychology, medical sociology, and digital infodemiology, and the MEDF is explicitly positioned relative to established health communication frameworks, including the i-frame and s-frame distinction (individual-level vs system-level intervention targets) and socioecological infodemic models, with each construct's novelty defined in relation to adjacent concepts in prior literature. The MEDF proposes that AI-driven health misinformation is distinctively dangerous due to its capacity to exploit variable individual receptivity to medical authority claims and to simultaneously lower epistemic thresholds across multiple trust layers. Population-level data indicate that individuals who frequently encounter health misinformation on social media are 1.66 times more likely to report systemic distrust of health care institutions (odds ratio 1.66, 95% CI 1.11-2.48). Perceptual studies document that listeners correctly identify AI-generated voice clones only about 60% of the time and perceive a cloned voice as identical to its real counterpart in approximately 80% of trials. Existing defenses—including Content Provenance and Authenticity standards, automated deepfake detection (showing area under the curve drops of up to 50% under real-world conditions), and prebunking interventions—are shown to address only subsets of the proposed cascade, leaving temporal and systemic layers substantially unmitigated. Four testable hypotheses are advanced for empirical validation. Addressing AI-driven health misinformation requires moving beyond individual-level i-frame interventions toward structural, s-frame policy responses calibrated to each layer of the MEDF cascade. Policymakers and platforms must implement source identity verification, clinician biometric protection protocols, cross-platform ecosystem governance, and proactive trust infrastructure, with particular urgency in lower- and middle-income country contexts where regulatory capacity and platform oversight are most limited.

JMIR Infodemiology 2026;6:e96664; doi: 10.2196/96664

Keywords: health misinformation; generative artificial intelligence; large language models; deepfakes; infodemiology; epistemic trust; health communication; conceptual framework; public health policy

Introduction: Epistemic Transformation and the Research Gap

The production and consumption of health information has historically been grounded in a robust institutional framework. Medical authority has been constructed through education, licensure, and professional accountability mechanisms, providing an epistemic assurance function for the credibility of health knowledge. The transformation of information ecosystems by algorithmic interfaces, accelerated by the proliferation of generative AI (GenAI), has brought this framework to a qualitative breaking point. The "infodemic" concept formulated by the World Health Organization during the COVID-19 pandemic foreshadowed this rupture; yet GenAI has fundamentally altered both the production logic and the scale of infodemics.

The existing literature largely addresses AI's role in health misinformation across two axes: threat vectors (hallucinations, deepfakes, and propaganda automation) and defensive opportunities (early warning systems, real-time verification, and health literacy). These axes are sometimes consolidated under the rubric of a "dual role," emphasizing that AI can both generate and detect misinformation through the same technical infrastructure. A recent scoping review theorized this paradox through the concept of "epistemic ambivalence"—AI's potential to simultaneously construct and erode public knowledge [1]. However, this framing leaves a fundamental question unresolved: why is AI-driven health misinformation qualitatively distinct from other forms of misinformation?

This viewpoint's central argument is as follows: the danger of AI in the health information ecosystem lies not in generating false information per se, but in its capacity to embed that false information simultaneously across multiple layers of human trust systems. This capacity operates through clinical language shielding (false claims wrapped in fluent medical terminology and fabricated citations), embodied authority transfer (deepfaked clinician faces and voices carrying claims their real counterparts never made), the synthetic chorus effect (the near-instantaneous fabrication of apparently independent corroborating sources), and structural epistemic erosion (the cumulative collapse of trust in medical institutions). Each layer amplifies the preceding one, producing cumulative harm that exceeds the sum of individual layers.

This framework is specific to the health domain. In political or commercial misinformation, audiences are conditioned to approach sources with some degree of skepticism. In health information, the opposite dynamic obtains: health information seekers frequently exhibit heightened receptivity to claims presented with apparent medical authority—a term preferred over more reductive framings of uniform vulnerability, in line with current National Academies of Sciences, Engineering, and Medicine guidance [2]. This receptivity is not homogeneous: empirical research documents meaningful variation in individual susceptibility across populations and contexts. Chronic disease patients show direct associations between health-related social media use and cyberchondria—excessive anxiety and false health beliefs mediated by negative affect [3]. Heightened receptivity is further documented among older adults [4], individuals experiencing depression [5], and populations experiencing psychological distress or loneliness [6]. While institutional trust in medical authority provides a powerful baseline orientation for health information seeking, this trust is neither universal nor culturally invariant: cross-cultural research documents substantial variation in public deference to medical authority, structured primarily by societal power distance and cultural egalitarianism [7,8]. Under the structural and cultural conditions where such institutional trust does operate, AI systematically targets and exploits it.

This paper is presented as a viewpoint and framework piece. The framework is a conceptual proposal that has not yet undergone empirical validation; testable hypotheses have been developed for each layer and future research agendas identified. The aim of this paper is to introduce the Multilayered Epistemic Disruption Framework (MEDF), which conceptualizes how AI-driven health misinformation structurally undermines public trust through four interdependent layers of cognitive and institutional disruption. Moving beyond binary threat–opportunity framing, the MEDF maps how generative technologies alter the structural conditions of health communication and derives layer-specific policy interventions and empirical research agendas. The structure is as follows: the Core Constructs section provides a glossary of specialized cross-disciplinary terms (Table 1) and formal definitions of the four layers; the Theoretical Framework section addresses conceptual positioning and construct originality; four subsequent sections elaborate each layer in detail; the paper then evaluates existing defense mechanisms, presents policy recommendations, and closes with testable hypotheses, conclusions, and limitations.

Table 1. Working definitions of specialized cross-disciplinary terms used in this paper.

Term	Working definition
Machine heuristic	The cognitive shortcut by which users automatically attribute objectivity and accuracy to machine-generated content [9].
Truth-default bias	The default tendency to presume that communication is honest unless suspicion is actively triggered, recently extended to human responses to generative AI content [10].
Liar's dividend	The benefit accruing to dishonest actors when public awareness of deepfakes allows authentic evidence to be dismissed as fabricated [11].

Term	Working definition
Epistemic trust	In the sense of Fonagy and colleagues: an individual's willingness to accept new communicated information as trustworthy, generalizable, and relevant to the self [12].
i-frame/s-frame distinction	The contrast between interventions targeting individual cognition and behavior (i-frame) and interventions targeting the systemic rules and structures within which individuals act (s-frame) [13].
Need for cognitive closure	The motivated desire to reach a firm answer and avoid ambiguity, leading individuals to seize on early explanations and then freeze on them [14].
Dual coding theory	The theory that information encoded simultaneously through verbal and visual channels leaves stronger memory traces than single-channel encoding [15].
Illusory truth effect	The increase in the perceived truth of a statement produced by repeated exposure, largely independent of source credibility [16].
Epistemic ambivalence	Generative AI's paradoxical capacity to simultaneously construct and erode public knowledge [1].
Astroturfing and coordinated inauthentic behavior	The manual orchestration of ostensibly independent accounts and messages to simulate organic grassroots consensus, typically for political or commercial ends.

Core Constructs: Definitions and Architecture of the MEDF

This viewpoint draws on specialized concepts from clinical medicine, public health, informatics, and communication research. Table 1 provides working definitions of these cross-disciplinary terms together with their source literature; the four MEDF constructs themselves are formally defined in the remainder of this section.

The MEDF is formally defined as an analytic framework that conceptualizes AI-driven health misinformation not as a content problem—discrete false claims to be detected and corrected—but as a structural process operating simultaneously across four distinct layers of the trust systems through which people assess health information. Its unit of analysis is the information ecosystem rather than the individual message or user, and its purpose is twofold: to differentiate four mechanisms of epistemic disruption that require different points of intervention, and to articulate a testable hypothesis about how these mechanisms compound one another. The MEDF is not an empirically validated causal model, nor a detection instrument; it is a conceptual lens whose propositions are advanced for empirical evaluation. The four constructs are formally defined here, before their theoretical positioning and elaboration in subsequent sections. They are ordered analytically from micro-level linguistic manipulation to macro-level institutional erosion; this ordering reflects the proposed cascade logic, not an established temporal sequence.

1. Discursive layer—clinical language shielding: the mechanism by which inaccurate, unproven, or fabricated health claims are embedded within sophisticated medical terminology, pseudoclinical reasoning, and structurally plausible but distorted or fabricated citations. The professional linguistic frame itself functions as a deceptive credibility signal, bypassing the user's content-level evaluation capacity.
2. Biometric layer—embodied authority transfer: the digital hijacking and weaponization of a verified medical professional's physical identity—including voice clones, facial expressions, and somatic cues via deepfakes—to disseminate health claims the clinician never authorized or made.
3. Temporal layer—the synthetic chorus effect: the automated, near-instantaneous creation of a dense, cross-channel information infrastructure—including clone websites, fabricated patient testimonials, synthetic social media profiles, and video content—centered around a single false claim, creating a persuasive illusion of independent consensus within hours.
4. Systemic layer—structural epistemic erosion: the macro-level degradation of public trust in real medical authorities and institutions, occurring as the pervasive presence of hyper-realistic health deepfakes produces cognitive fatigue and chronic skepticism toward authentic health communication—what Chesney and Citron [11] termed the liar's dividend.

The structural logic of the MEDF posits a sequential cascade: clinical language shielding (layer 1) is hypothesized to lower baseline cognitive resistance, rendering the user more susceptible to biometric manipulation (layer 2), which is reinforced by fabricated consensus (layer 3), ultimately contributing to systemic institutional distrust (layer 4). This cascade, if empirically confirmed, would produce harm that is nonadditive. Each layer can cause harm independently; the cascade logic constitutes the framework's most contestable and most important empirical claim, elaborated in the Interlayer Relationships section below. Figure 1 illustrates the framework's structure, the sequential cascade between layers, and the cumulative reinforcement dynamic. Each layer is hypothesized to lower the user's resistance to the next (solid arrows), while their combined operation produces nonlinear, cumulative disruption (dashed reinforcement). Table 2 differentiates the four constructs by their primary trust target, core mechanism, theoretical basis, and intervention point.

Figure 1. The Multilayered Epistemic Disruption Framework (MEDF): a proposed cascade across four interdependent layers of health-information trust. The cascade is a theoretical proposition advanced for empirical testing and is not yet experimentally validated.

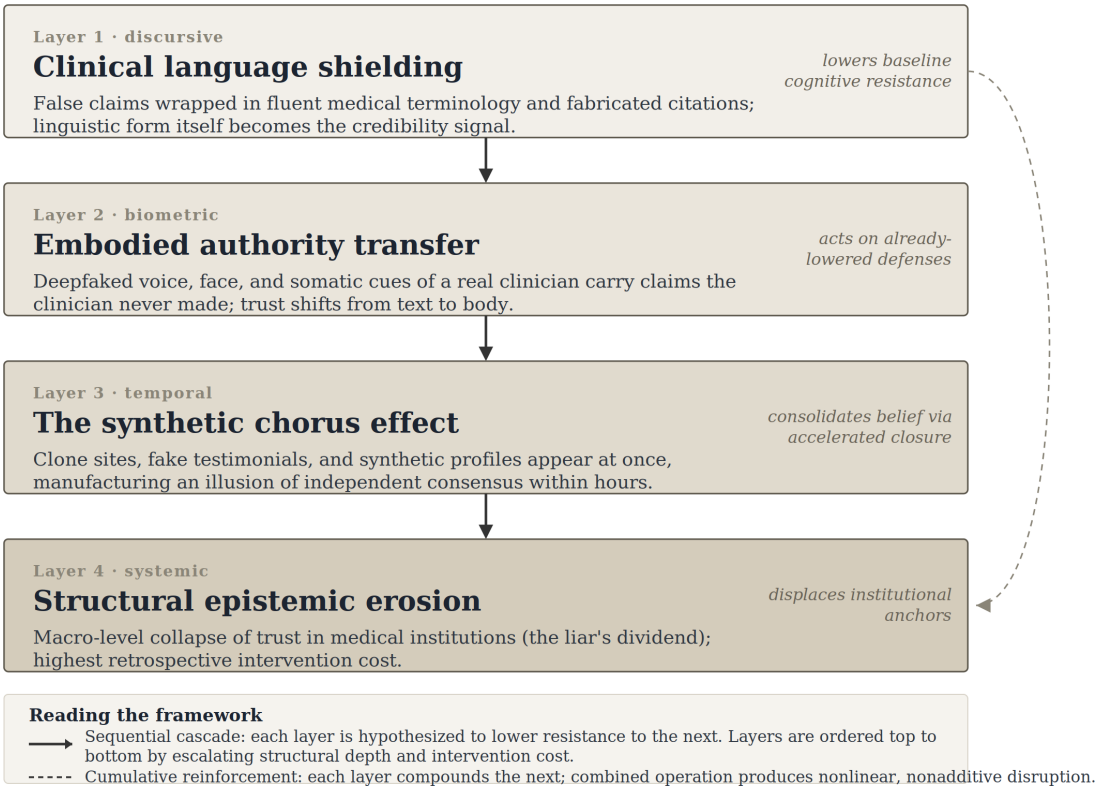

Table 2. Differentiation of the four MEDF[a] constructs by trust target, mechanism, theoretical basis, and intervention point.

Layer and construct	Primary trust target	Core mechanism	Principal theoretical basis	Primary intervention point
Layer 1: discursive: clinical language shielding	Linguistic markers of clinical competence	False claims embedded in fluent medical terminology and fabricated citations	Machine heuristic [9]; truth-default bias [10]	Source identity verification
Layer 2: biometric: embodied authority transfer	Embodied identity of recognized clinicians	Deepfaked face and voice carry claims the clinician never made	Deepfake perception research [17]	Clinician identity protection
Layer 3: temporal: synthetic chorus effect	Perceived independence and consensus of sources	Near-instantaneous multi-channel fabrication of corroborating sources	Source multiplicity [16, 18]; need for cognitive closure [14]	Ecosystem-level governance
Layer 4: systemic: structural epistemic erosion	Institutional credibility of medical authority	Pervasive synthetic content induces chronic skepticism toward authentic communication	Liar's dividend [11]; posttruth distrust dynamics [19]	Proactive trust infrastructure

[a]MEDF: Multilayered Epistemic Disruption Framework.

Theoretical Framework and Conceptual Positioning

Gap in the Literature and Conceptual Positioning

The MEDF is distinguished from prior work by 3 features: it treats threat and defense as two faces of one mechanism rather than separate lists; it defines the layers as independent but interconnected, such that simultaneous operation yields nonlinear rather than additive harm; and it is domain-specific. Layered frameworks exist in general disinformation research, but none systematically theorizes the vulnerability dynamics particular to health—the clinical language shield, the lowered epistemic threshold that accompanies patient anxiety, and the reflex of deference to clinical authority.

The closest conceptual neighbors in the literature are as follows:

1. Sundar's machine heuristic [9] describes the automatic trust users extend to machine-generated content and supplies the psychological foundation for clinical language shielding. The fourth layer builds on the liar's dividend [11]. The epistemic trust literature by Fonagy et al [12] provides grounding for understanding trust mechanisms in health contexts. Park and Nan's [1] epistemic ambivalence offers the first systematic theorization of AI's paradoxical capacity to both construct and erode public knowledge. These works are the MEDF's predecessors; none of them,

however, integrates these mechanisms into a unified framework specific to health misinformation.

2. A word on conceptual boundaries: clinical language shielding differs from the "generative illusion" [20] and "authority cue" concepts in that those capture general large language model (LLM) competence illusions, while clinical language shielding addresses the health-specific version of that illusion and its interaction with patient vulnerability. Likewise, the synthetic chorus effect differs from astroturfing and coordinated inauthentic behavior in that those concepts rest on manual coordination and political motivation; the synthetic chorus effect describes a low-cost, near-instantaneous multichannel production capacity now accessible to individual actors, and its specific consequences in health contexts.
3. Positioning within existing frameworks: a foundational distinction in recent behavioral public policy and health communication research is the contrast between the i-frame and the s-frame [13]. The i-frame locates solutions at the level of individual cognition and behavior; the s-frame addresses the structural, systemic conditions that shape an information ecosystem. Traditional interventions against health misinformation—individual media literacy campaigns, localized fact-checking, and content labeling—operate within the i-frame. The MEDF is explicitly an s-frame model: its central argument is that the primary harm of GenAI does not stem from individual users being deceived case by case, but from the systematic restructuring of the digital environment in ways that lower epistemic thresholds for all users simultaneously.
4. Established public health frameworks for infodemic management provide valuable structural precedents. The 4-I Fact Framework [21] offers a structured approach to information assessment and institutional trust rebuilding. Socioecological and environmental health models [22] have mapped how misinformation circulates at micro, meso, and macro levels. These frameworks were, however, largely developed for pre-GenAI disinformation dynamics or manual, human-coordinated campaigns. The MEDF builds on their structural orientation while addressing three specific dynamics outside their scope: ecosystem-level automation at near-zero marginal cost; the weaponization of clinical epistemology; and a nonlinear cascade logic in which each layer actively degrades resistance to subsequent layers.
5. From structural approaches to epistemic injustice [23], the MEDF draws the insight that epistemic harm operates not only through individual deception but through the systematic undermining of the institutional structures through which credibility is assessed and granted. Applied to AI-mediated health communication, what is eroded is not merely individual belief, but the structural conditions under which clinical authority can function as a credibility signal at all.
6. Construct originality: each of the MEDF's four constructs requires explicit positioning relative to adjacent concepts in prior literature. The locus of the framework's novelty claim should first be stated precisely: although the constructs vary in their distance from existing concepts, the MEDF's primary contribution resides not in any single component but in the interaction architecture—the integration of four mechanisms into a unified, health-specific structure with a hypothesized compounding logic and a layer-specific intervention mapping. Within that architecture, the individual constructs are positioned as follows. Clinical language shielding is the framework's most original analytical contribution. While analogous dynamics—medical framing as a persuasion strategy, professional language as a credibility heuristic—have been documented in health communication and contested illness research, prior literature addresses these as human-authored, limited-scale phenomena. Clinical language shielding formalizes the AI-specific version, characterized by LLM fluency, scalability, and the automated generation of fabricated clinical citations. Embodied authority transfer draws on established deepfake research; its specific contribution is formalizing biometric identity theft as a proposed public health mechanism, distinct from privacy violation or political disinformation contexts, and treating the empirical confirmation of this mechanism as an explicit research agenda item. The synthetic chorus effect is theoretically novel in combining source multiplicity, simultaneous cross-channel presence, and GenAI-enabled instantaneous deployment; existing astroturfing and coordinated inauthentic behavior frameworks, designed for manual, slower-moving operations, do not capture this dynamic. Structural epistemic erosion draws directly on the liar's dividend [11]; the MEDF's contribution is applying this concept to health communication and articulating its differential geographic and institutional impact.

Interlayer Relationships: The Cumulative Disruption Dynamic

The question of how the four layers reinforce one another is the framework's most critical, and most contestable dimension. A fair critique will note that layers overlap in practice: clinical language shielding operates inside deepfake videos; the synthetic chorus effect directly feeds structural erosion. This overlap is not a methodological problem; it is the framework's central claim. The layers intersect. Their analytical value lies precisely in the fact that they require different points of intervention. Countering clinical language shielding calls for source authentication; countering embodied authority transfer calls for identity protection; countering the synthetic chorus effect requires ecosystem-level governance; and countering structural epistemic erosion requires proactive trust infrastructure. No single intervention can address all four simultaneously, and that is why treating AI-driven health misinformation as a single problem with a single fix has so far produced limited results.

The theoretical basis for the nonlinear, cumulative character of this harm rests on two bodies of evidence already

present in the literature. First, the work by Fonagy et al [12] on epistemic trust in psychotherapeutic contexts established that trust mechanisms are hierarchically organized, with credibility assessments at one level modulating openness to information at the next. It must be stated transparently that this framework was developed to account for mentalizing processes in clinical relationships; applying it to AI-mediated health information ecosystems is a deliberate theoretical extension, and the cascade mechanism proposed here should be read as a theoretically motivated hypothesis rather than an established finding, pending direct experimental validation. The proposed dynamic is as follows: when clinical language shielding successfully mimics technical competence, it is hypothesized to lower the user's baseline cognitive resistance; deepfake imagery then encounters a subject whose epistemic defenses have already been partially lowered; the synthetic chorus effect consolidates a belief formed under conditions of reduced resistance; and structural erosion sets in as those beliefs begin to displace institutional anchors. This proposed cascade, if empirically confirmed, would produce harm that is nonadditive.

Second, Kruglanski and Webster's [14] need for cognitive closure establishes that once a satisfying explanation is reached, the tendency is to defend rather than reopen it. Critically, the speed at which closure is reached increases with the number of consistent signals encountered. Multichannel exposure, the defining feature of the third layer, does not merely add to the probability of belief; it accelerates the closure process in a nonlinear fashion, because each additional consistent signal reduces the cognitive resources the subject is willing to invest in continued scrutiny. The synthetic chorus effect exploits this dynamic directly. Taken together, these two frameworks supply the psychological mechanism that bridges individual layer effects to the cumulative harm the framework predicts: prior trust erosion lowers the threshold for closure, and closure accelerates with signal density in a way that no linear model can capture.

One clarification prevents misreading: the cascade is not advanced as a claim of necessary sequence. Any layer can be activated independently: a user may encounter a clinician deepfake without prior exposure to LLM-generated text, and structural erosion can be fed by the mere public awareness that synthetic content exists [11]. The framework's claim is that when layers cooccur, their joint operation is hypothesized to produce more-than-additive harm; it is the compounding, not the ordering, that constitutes the central empirical proposition.

Layer 1—Discursive: Clinical Language Shielding

Mechanism and Positioning in the Literature

LLMs can command virtually any domain of expertise at high fluency, including medical terminology. That fluency, combined with the models' inability to verify the accuracy of what they produce, generates a health-specific threat: clinical language shielding. The mechanism works as follows: a false health claim is wrapped in medical terminology, plausible-sounding clinical reasoning, and fabricated or distorted references. When a user reads the result, it is not the content they evaluate, it is the language itself that functions as the credibility signal. This explains both Sundar's machine heuristic [9]—the automatic trust attribution users extend to machine-generated output—and the paradox identified by Markowitz and Hancock [10]: LLMs are more truth-biased than humans, which is precisely what makes their production of false claims so dangerous.

Empirical Evidence

A 2026 Lancet Digital Health study by Omar and colleagues provided systematic documentation of this mechanism, analyzing 3.4 million prompts across 20 LLMs [24]. Testing with real hospital discharge notes, social media health myths, and 300 clinician-approved scenarios, the researchers found that existing safety filters could not reliably distinguish true from fabricated claims when those claims were wrapped in familiar clinical language. A related randomized controlled experiment involving medical students [25] found that misleading AI explanations significantly degraded diagnostic accuracy, while accurate AI explanations produced no statistically significant improvement. This asymmetry, the harm of false information outweighing the benefit of true information, is precisely what makes clinical language shielding so dangerous in health contexts.

The scope of this layer extends well beyond supplement marketing. Research on vaccine hesitancy shows that LLMs can present inaccurate claims about vaccine safety within a scientific-looking frame when safety guardrails are bypassed [26]. In oncology contexts, herbal treatments and off-protocol regimens can be legitimized through clinical terminology. The same system-instruction vulnerabilities documented by Modi et al [26] make it technically feasible to frame unproven treatments in scientifically credible language once guardrails are circumvented.

Testable Hypothesis

Hypothesis 1 predicts that identical false claims framed in clinical terminology produce higher belief and compliance than in everyday language.

Layer 2—Biometric: Embodied Authority Transfer

Mechanism

Where clinical language shielding operates at the discursive level, deepfake technology carries the threat into a biometric one: it is no longer language that provides reassurance, but the body itself. A recognized physician's face, voice, posture, and eye contact activate social recognition and parasocial trust mechanisms that operate below conscious scrutiny. This framework terms this dynamic embodied authority transfer: the weaponization of a real medical authority's biometric

identity to carry claims that authority never made. Barrington et al [17] found that people cannot reliably distinguish AI-generated voice clones from real speech: participants perceived a cloned voice to be the same as its real counterpart approximately 80% of the time, and correctly identified a voice as AI-generated only about 60% of the time. Liu et al [27] demonstrated experimentally that AI doctor avatars appearing older elicit higher trust ratings, suggesting that deepfake designers are deliberately modeling preexisting cognitive biases in their audiences.

Documented Patterns and Topic Distribution

A Bitdefender Labs investigation covering March–May 2024 on Meta platforms identified over 1000 deepfake health videos promoting more than 40 supplement products, targeting diabetes, joint pain, vision problems, and cardiac conditions [28]. A Full Fact investigation documented networks of accounts using deepfaked academics and clinicians across TikTok (ByteDance) and Instagram (Meta Platforms, Inc) [29], and reporting in the medical press has described the same impersonation pattern across other major platforms [30]. A 2026 statement by the American Medical Association chief executive officer in STAT News [30] documented the same pattern concentrated around glucagon-like peptide-1 alternatives, weight loss pills, and diabetes treatments.

The topics targeted are not random. Diabetes, obesity, and weight management, menopause, joint pain, and antiaging products cluster around conditions characterized by chronicity, stigma, waiting-list pressure, and a perceived inadequacy of conventional care. This mirrors broader health misinformation patterns: vaccine misinformation research shows an identical vulnerability mapping, with high-anxiety and low-institutional-trust contexts as primary targets [31]. The Wellness Nest case illustrates the dynamic concretely: Professor David Taylor-Robinson's expertise in child health was weaponized to promote menopause supplements — a deliberate mismatch between domain authority and topic that reveals how the mechanism works. Affected clinicians have had to spend hours trying to have deepfake content removed, defending both their professional reputation and their patients' safety.

Testable Hypothesis

Hypothesis 2 predicts belief and compliance increase from text to anonymous avatar to recognized-clinician deepfake, with the largest gap at the deepfake step.

Layer 3—Temporal: The Synthetic Chorus Effect

Mechanism

The synthetic chorus effect names a specific kind of skepticism erosion: the dissolution of doubt that occurs when AI constructs a comprehensive credibility infrastructure around a single claim—websites, academic-looking articles, patient testimonials, clinician blogs, social media profiles, and videos—simultaneously and within hours. This mechanism is best understood as the intersection of three well-established psychological processes, now operating in an environment their original theorists could not have anticipated. First, source multiplicity research consistently shows that information attributed to multiple independent sources is perceived as substantially more credible than the same information from a single source [16,18]. Second, repeated exposure, independent of source, increases the perceived truth of a statement: the illusory truth effect [16]. The synthetic chorus effect is analytically distinct from, and complementary to this mechanism: the illusory truth effect operates through temporal repetition of the same claim across successive encounters, whereas the chorus operates through the structural density of apparently independent corroborating sources encountered at a single moment. Repetition builds familiarity over time; the chorus manufactures consensus all at once. Third, Paivio's [15] dual coding theory and Mayer's [32] cognitive theory of multimedia learning demonstrate that information encoded simultaneously through verbal and audiovisual channels leaves a stronger, more durable memory trace than single-channel exposure, reinforcing both retention and subjective certainty. Add to this Kruglanski and Webster's [14] need for cognitive closure: once a person reaches a satisfying explanation, they tend to defend that judgment rather than reopen it. Multichannel exposure accelerates that closure. The synthetic chorus effect exploits all of this at once. Every apparent source—the website, the blog, the patient story, and the video—was produced by a single actor within hours. The user perceives independent corroboration; what they are experiencing is a coordinated illusion of consensus.

The Temporal Factor and Collapse of Defensive Capacity

The most critical feature of this layer is its speed. Traditional disinformation operations took time to build credibility: the age of a website, the history of a social media account, accumulated engagement patterns—these were all verification signals. AI eliminates that cost. A coherent, mutually reinforcing information ecosystem spanning multiple channels can now be built in hours. Even where health information consumers approach isolated videos or single promotional pages with skepticism, the framework proposes that encountering a whole system—website, blog, social media profiles, patient testimonials, and user reviews—can erode that skepticism considerably; this proposition is formalized in hypothesis 3 below.

The synthetic chorus effect's emergence is best understood against the backdrop of how social media had already transformed information environments before GenAI. Research on online news spread established that false stories diffuse faster, more broadly, and more deeply than true ones across social networks, driven primarily by human sharing behavior [33]. Earlier analyses documented how the structural incentive architecture of social media, algorithmic amplification of engagement-maximizing content, had

already converted passive information consumers into active amplifiers, a dynamic Phillips [34] described as the oxygen of amplification in the context of political disinformation. GenAI did not create this vulnerability; it automated its exploitation at a previously inaccessible scale. The concepts of astroturfing and coordinated inauthentic behavior describe the political predecessors of the synthetic chorus effect, but were developed around the manual coordination capacity of the pre-GenAI era. The DC Weekly operation's documented increase in content production through AI use [35] illustrates the scale shift: what previously required organized human networks can now be achieved by a single actor within hours.

Empirical Gap and Testable Hypothesis

No empirical study has directly tested the synthetic chorus effect—this is one of the paper's most important research gap findings. Existing disinformation research predominantly addresses isolated content units. Controlled experiments measuring the effect of simultaneous multichannel presence on user trust and health behavior represent the most pressing research agenda this framework generates.

Hypothesis 3 predicts simultaneous multichannel presence raises trust nonlinearly, the gain from two to four channels exceeding that from one to two.

Layer 4—Systemic: Structural Epistemic Erosion

Mechanism: The Liar's Dividend and Institutional Trust

The cumulative effect of the first three layers feeds the fourth and deepest: structural epistemic erosion. This dynamic, what Chesney and Citron [11] named the liar's dividend, the way that the mere existence of fake content erodes trust in real content, is particularly destructive in health communication. A patient who has watched a deepfake of a doctor may approach their next real clinical consultation, or the next message from a public health authority, with a residual skepticism that was not there before. This erosion does not target individual health decisions; it undermines institutional trust in medical authority as such. A nationally representative US study of 5041 adults found that those who reported encountering substantial health misinformation on social media were 1.66 times more likely to report low trust in the health care system (odds ratio 1.66, 95% CI 1.11-2.48) [36].

Documented cases, including patients who abandoned prescribed treatments after watching deepfake physician endorsements of alternative remedies [29,30], illustrate how medical authority can be structurally displaced. Health communication research shows that once institutional distrust is established, the effectiveness of corrective messages drops substantially [19]. On this proposed logic, the fourth layer may represent the most consequential: structural epistemic erosion carries the highest projected retrospective intervention cost.

Epistemic Fatigue and Beyond the Global North

A second mechanism within structural erosion is epistemic fatigue: the cognitive exhaustion of constantly asking which information is real and which was generated by AI. Research shows that exposure to highly realistic disinformation weakens confidence in distinguishing truth from fabrication—a kind of cynicism some researchers have begun calling truth fatigue [37]. The authority being eroded is, as medical sociology has long documented, a historically constructed achievement rather than a natural fact: professional dominance over health knowledge and the lay–expert boundary were institutionalized over more than a century of professionalization [38,39]. Structural epistemic erosion can therefore be read as the synthetic acceleration of that construction's reversal—an authority assembled over a century, destabilized at the pace of automated content generation. Experimental evidence is consistent with this proposed dynamic: exposure to deepfakes has been shown to generate uncertainty more readily than outright deception, and that uncertainty in turn reduces trust in news encountered on social media [40] — a cynicism pathway that structural epistemic erosion extends to the health domain.

The geographic dimension of this layer deserves attention. Countries with developed digital infrastructure but relatively weak regulatory frameworks and uneven health literacy, for example, Brazil, India, and Nigeria, represent the contexts where this framework's predictions are likely sharpest. In lower- and middle-income countries (LMICs), deepfake detection infrastructure is limited, platform oversight is thin, and clinicians have limited capacity to monitor their own digital footprint. The MEDF's health-domain specificity is most pronounced in these settings: where baseline institutional health distrust is already high, structural epistemic erosion can advance considerably faster.

Testable Hypothesis

Hypothesis 4 predicts cumulative deepfake exposure reduces trust in clinicians and institutions and depresses health-seeking behavior, most sharply where baseline distrust is high.

Evaluation of Existing Defense Mechanisms

Content Labeling and Provenance Standards

China's Cyberspace Administration regulations, effective September 2025, require all AI-generated content to carry labeling at both the visual and metadata level. The EU AI Act and US state-level legislation impose comparable transparency obligations. Technical standards such as C2PA (Coalition for Content Provenance and Authenticity) aim for cryptographic verification of content origin. These measures offer a partial response to layers 1 and 2, but carry two fundamental limitations.

The first is an enforcement gap: synthetic health content frequently circulates without any transparency label, as metadata can be stripped during peer-to-peer reposting and visual watermarks can be deprioritized in platform interface designs. Initiatives such as Google's SynthID and TikTok's C2PA integration are promising, but their effectiveness depends on platform compliance that is not yet mandatory or consistently monitored. The second limitation is that labeling does not directly target the synthetic chorus effect (Layer 3) or structural epistemic erosion (Layer 4): flagging a single content unit within a coherent ecosystem does not undo the system's persuasive force.

Detection Systems and Their Limitations

The World Health Organization's EARS (early AI-supported response with social listening) platform and similar systems hold real potential for early detection of health misinformation. However, the Deepfake-Eval-2024 benchmark found that state-of-the-art deepfake detection models show area under the curve drops of up to 50% for video, 48% for audio, and 45% for images when evaluated on real-world content versus controlled datasets [41]. The synthetic chorus effect compounds the challenge further: when campaign components are distributed across platforms and appear independent of one another, algorithmic identification of their coordination is technically difficult.

Prebunking: The Individual-Level Limitation

The demonstrated inadequacy of post hoc correction has driven interest in prebunking. Experimental evidence supports the effectiveness of gamified prebunking interventions in certain contexts [42]. Viewed through the MEDF lens, however, prebunking carries a structural limitation that the i-frame and s-frame distinction [13] makes precise: it is an i-frame intervention targeting individual cognition, whereas the synthetic chorus effect operates at the s-frame level of the information environment itself. Equipping individuals to critically evaluate isolated content units may not be enough to hold back an ecosystem-level manipulation. A person who remains skeptical of a single piece of content can still find their defenses eroded by a coherent ecosystem of dozens of apparently independent sources. Individual media literacy is necessary but cannot substitute for structural, s-frame interventions at the ecosystem level.

Policy Recommendations

The MEDF makes a straightforward argument: because each of its four layers requires a different point of intervention, policy frameworks that address only one or two of them will leave significant damage pathways open. Consistent with the framework's s-frame orientation [13], the recommendations below target the structural conditions of the information ecosystem rather than individual user behavior. They are derived directly from the framework's structure and are calibrated against what existing initiatives address and what they miss.

Source Identity Verification

An effective defense against clinical language shielding must target the source, not the language. This means developing mandatory source authentication standards for AI systems that generate health information, and expanding the availability of health communication AI models trained on peer-reviewed medical literature with strengthened safety barriers. The practical obstacles are real: platform-level compliance requires global coordination, and companies such as Meta and Google may resist such constraints as encroachments on commercial autonomy. Joint American Medical Association–World Health Organization platform standardization efforts are promising in this space but have not yet produced binding frameworks.

Clinician Identity Protection Protocols

Recognizing clinician identity as a legally protected professional asset, as the American Medical Association has proposed, and requiring platforms to verify accounts using physician identities for health content are critical steps. The US TAKE IT DOWN Act (May 2025) mandates removal of synthetic personal content within 48 hours, but targets privacy violations rather than health misinformation. A health-specific equivalent does not yet exist. Article 50 transparency requirements in the European Union AI Act could partially close this gap, but enforcement mechanisms need strengthening. A further structural constraint follows from the framework's own temporal logic: false content diffuses significantly faster than true content across social networks [33], and the synthetic chorus effect compresses ecosystem construction into hours, so a 48-hour removal window can intervene only after the bulk of organic dissemination has already occurred. Removal mandates are necessary, but they are temporally outpaced by the threat they target.

Ecosystem-Level Governance

The synthetic chorus effect requires interventions that go beyond individual content moderation. Risk-scoring systems that evaluate account age, engagement patterns, and content consistency together, and shared threat intelligence mechanisms that enable detection of coordinated campaigns across platforms, are the natural policy response. Researcher–platform collaboration models such as Stanford Internet Observatory's Spamouflage tracking offer good precedents; health-specific equivalents have not yet been institutionalized. Cross-platform threat intelligence also presupposes moderation capacity that is unevenly distributed: in the LMIC contexts where this framework predicts the sharpest effects, platform moderation resources and local-language coverage are thinnest, and operations hosted outside national jurisdictions remain largely beyond the reach of instruments such as the European Union AI Act.

Proactive Trust Infrastructure

Structural epistemic erosion cannot be meaningfully addressed through reactive correction alone. User-friendly tools that allow clinicians to cryptographically sign authentic content (C2PA-compatible), institutional mechanisms

for health organizations to mark their digital presence as verifiable, and health literacy programs redesigned to include ecosystem awareness, teaching people to recognize coordinated information systems, not just individual false claims, all belong in this layer. In LMIC contexts, building this infrastructure requires additional resources; it also offers the greatest potential impact in precisely those settings. Given the temporal limits of removal noted above, platform architectures could also explore correction mechanisms that preserve rather than delete. Because removal can feed a censorship narrative consistent with the liar's dividend [11], and because corrective messages lose effectiveness once institutional distrust is established [19], replacing an identified synthetic impersonation with a clinician-authorized disclaimer, while leaving its distribution links intact, would convert the falsehood's accumulated reach into a delivery vector for authentic correction. Such interception protocols remain untested and would face substantial regulatory, privacy, and platform-liability constraints; they are advanced here as a direction for design research rather than as deployable solutions.

A final consideration concerns sequencing and evaluability. These four intervention classes differ in their deployment horizons. Provenance labeling and ecosystem-level risk governance can be pursued immediately through existing instruments, C2PA standards and, in the European Union, the systemic-risk assessment obligations that the Digital Services Act already imposes on very large online platforms, whereas clinician biometric protection requires new institutional mechanisms, and proactive trust infrastructure is a long-horizon investment whose effects compound slowly. The framework's hypotheses double as evaluation criteria: the relationship that hypothesis 4 predicts between cumulative deepfake exposure and institutional trust, for example, provides a measurable baseline against which the effectiveness of any of these interventions can be assessed over time. The recommendations are thus offered not as a finished regulatory program but as a layer-indexed agenda whose components can be prioritized by deployment horizon and monitored through the same empirical apparatus the framework proposes.

Testable Hypotheses, Conclusions, and Limitations

Research Agenda

Four testable hypotheses, formally stated in their respective layer sections above, are proposed to guide empirical validation of the MEDF. In summary: hypothesis 1 predicts a clinical-terminology advantage in belief and behavioral compliance for identical false claims (layer 1); hypothesis 2 predicts an ordered increase in trust from text to anonymous synthetic avatar to recognized-clinician deepfake, with the largest gap at the deepfake step (layer 2); hypothesis 3 predicts nonlinear gains in trust under simultaneous multichannel presence (layer 3); and hypothesis 4 predicts that cumulative deepfake exposure erodes trust in real clinicians, health institutions, and health-seeking behavior, most sharply in populations with high baseline institutional distrust (layer 4).

A and B experiments are the natural methodology for testing hypothesis 1, hypothesis 2, and hypothesis 3; longitudinal survey designs for hypothesis 4; and cross-cultural comparative studies for the LMIC dimensions of hypothesis 3 and hypothesis 4. Full operational specifications for all four hypotheses are provided in Multimedia Appendix 1.

Conclusions

This viewpoint has proposed an analytical model, the MEDF, that addresses AI's role in health misinformation beyond the dual-role threat-opportunity framing that currently dominates the literature. The argument is that the danger does not lie in the falsehood of individual content units but in AI's capacity to embed false information simultaneously across four distinct layers of human trust systems.

The paper makes three contributions. First, it systematically theorizes the vulnerability dynamics specific to health misinformation, clinical language shielding, embodied authority transfer, the synthetic chorus effect, and structural epistemic erosion, and explicitly differentiates each from similar concepts in adjacent literature. Second, it maps which existing defense mechanisms address which layers and which layers they leave untouched, deriving policy recommendations with an honest account of their practical constraints. Third, it acknowledges the research gap openly, advances testable hypotheses for each layer, and contributes to shaping an empirical research agenda.

Limitations

Three limitations deserve acknowledgment. First, the MEDF is an untested theoretical framework. The synthetic chorus effect in particular, and the cumulative harm claim more broadly, are propositions awaiting controlled experimental validation. Second, the paper draws predominantly on English-language and Western-context literature. Different authority perceptions, local language dynamics, and digital literacy patterns in LMIC contexts are underrepresented—a limitation that has been flagged throughout the analysis but that comparative research will need to systematically address. Third, the documented cases come predominantly from the supplement and commercial health fraud context; systematic mapping of similar patterns in vaccine misinformation, pandemic management, and oncology would substantially extend the framework's scope.

Acknowledgments

In accordance with JMIR Publications policy on the use of artificial intelligence (AI) and large language models (LLMs) in scholarly work, the author discloses the following. The intellectual core of this manuscript—the conception of the Multilayered Epistemic Disruption Framework (MEDF), the definition and differentiation of its four constructs, the cascade logic linking

them, the policy analysis, and the testable hypotheses—was developed by the human author, who originally drafted the work in Turkish. LLM-based tools (OpenAI ChatGPT, Google Gemini, and Anthropic Claude) were used as assistive instruments under the author's direction for the following tasks: translation of the author's Turkish-language drafts into academic English; restructuring of the manuscript to conform to JMIR formatting requirements (eg, the numbered reference style); language editing for concision and readability; and assistance in locating, formatting, and verifying bibliographic references during revision. All AI-assisted output was reviewed, edited, and verified by the author, who takes full responsibility for the accuracy, integrity, and originality of the entire manuscript, including all claims, citations, and conclusions. The author confirms that no AI tool is listed or qualifies as an author, consistent with International Committee of Medical Journal Editors and JMIR authorship criteria.

Funding

No specific funding was received for the research. The article processing charge was partially supported by İstanbul Medipol University.

Data Availability

Not applicable. This is a viewpoint paper presenting a theoretical framework; no datasets were generated or analyzed.

Authors' Contributions

MM, as sole author, conceived the framework and all of its constructs, conducted the literature review, drafted the manuscript in Turkish, directed and verified the AI-assisted translation and formatting process described in the disclosure above, and reviewed and approved the final English text.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Full operationalization of the four MEDF hypotheses, including proposed conditions, predicted direction of effect, and indicative empirical designs for each layer.
[DOCX File (Microsoft Word File), 12 KB-Multimedia Appendix 1]

References

1. Park S, Nan X. Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. AI & Soc. Feb 2026;41(2):1501-1515. [doi: 10.1007/s00146-025-02620-3]
2. National Academies of Sciences, Engineering, and Medicine. Understanding and addressing misinformation about science. The National Academies Press; 2024. [doi: 10.17226/27894]
3. Lai YK, Lai Z, Zhao X. Counteracting cyberchondria in Chinese chronic disease patients: The divergent roles of health-related social media use and online patient-centered communication. Patient Educ Couns. Dec 2025;141:109337. [doi: 10.1016/j.pec.2025.109337] [Medline: 40945192]
4. Lyons B, King AJ, Barter RL, Kaphingst KA. Exposure to low-credibility online health content is limited and is concentrated among older adults. Nat Aging. Feb 2026;6(2):454-462. [doi: 10.1038/s43587-025-01059-x] [Medline: 41639217]
5. Delmastro M, Paciello M. Depression, reduced education, and bias perceptions as risk factors of beliefs in misinformation. Sci Rep. Sep 30, 2022;12(1):16408. [doi: 10.1038/s41598-022-20640-7] [Medline: 36180772]
6. Vidgen B, Taylor H, Pantazi M, Anastasiou Z, Inkster B, Margetts H. Understanding Vulnerability to Online Misinformation. The Alan Turing Institute; 2021. URL: https://www.turing.ac.uk/news/publications/understanding-vulnerability-online-misinformation [Accessed 2026-07-27]
7. Borg MA, Waisfisz B, Frank U. Quantitative assessment of organizational culture within hospitals and its relevance to infection prevention and control strategies. J Hosp Infect. May 2015;90(1):75-77. [doi: 10.1016/j.jhin.2014.12.015] [Medline: 25676113]
8. Touboul-Lundgren P, Jensen S, Drai J, Lindbæk M. Identification of cultural determinants of antibiotic use cited in primary care in Europe: a mixed research synthesis study of integrated design "Culture is all around us". BMC Public Health. Sep 17, 2015;15:908. [doi: 10.1186/s12889-015-2254-8] [Medline: 26381376]
9. Sundar SS. The MAIN model: a heuristic approach to understanding technology effects on credibility. In: Metzger MJ, Flanagin AJ, editors. Digital Media, Youth, and Credibility. MIT Press; 2008:73-100. URL: https://search.issuelab.org/resources/875/875.pdf [Accessed 2026-07-27]
10. Markowitz DM, Hancock JT. Generative AI are more truth-biased than humans: a replication and extension of core truth-default theory principles. J Lang Soc Psychol. Mar 2024;43(2):261-267. [doi: 10.1177/0261927X231220404]

11. Chesney R, Citron DK. Deep fakes: a looming challenge for privacy, democracy, and national security. Calif Law Rev. 2019;107:1753-1819. [doi: 10.15779/Z38RV0D15J]
12. Fonagy P, Luyten P, Allison E, Campbell C. Mentalizing, epistemic trust and the phenomenology of psychotherapy. Psychopathology. 2019;52(2):94-103. [doi: 10.1159/000501526] [Medline: 31362289]
13. Chater N, Loewenstein G. The i-frame and the s-frame: how focusing on individual-level solutions has led behavioral public policy astray. Behav Brain Sci. Sep 5, 2022;46:e147. [doi: 10.1017/S0140525X22002023] [Medline: 36059098]
14. Kruglanski AW, Webster DM. Motivated closing of the mind: "seizing" and "freezing". Psychol Rev. Apr 1996;103(2):263-283. [doi: 10.1037/0033-295x.103.2.263] [Medline: 8637961]
15. Paivio A. Mental Representations: A Dual Coding Approach. Oxford University Press; 1986. [doi: 10.1093/acprof:oso/9780195066661.001.0001]
16. Nadarevic L, Reber R, Helmecke AJ, Köse D. Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. Cogn Res Princ Implic. Nov 11, 2020;5(1):56. [doi: 10.1186/s41235-020-00251-4] [Medline: 33175284]
17. Barrington S, Cooper EA, Farid H. People are poorly equipped to detect AI-powered voice clones. Sci Rep. Mar 31, 2025;15(1):11004. [doi: 10.1038/s41598-025-94170-3] [Medline: 40164656]
18. Harkins SG, Petty RE. Information utility and the multiple source effect. J Pers Soc Psychol. 1987;52(2):260-268. [doi: 10.1037/0022-3514.52.2.260]
19. Lewandowsky S, Ecker UKH, Cook J. Beyond misinformation: understanding and coping with the "post-truth" era. J Appl Res Mem Cogn. 2017;6(4):353-369. [doi: 10.1016/j.jarmac.2017.07.008]
20. Joseph J, Jose B, Jose J. The generative illusion: how ChatGPT-like AI tools could reinforce misinformation and mistrust in public health communication. Front Public Health. 2025;13:1683498. [doi: 10.3389/fpubh.2025.1683498] [Medline: 41080885]
21. Sundelson AE, Jamison AM, Huhn N, Pasquino SL, Sell TK. Fighting the infodemic: the 4 i Framework for Advancing Communication and Trust. BMC Public Health. Aug 30, 2023;23(1):1662. [doi: 10.1186/s12889-023-16612-9] [Medline: 37644563]
22. John JN, Gorman S, Scales D. Understanding interventions to address infodemics through epidemiological, socioecological, and environmental health models: framework analysis. JMIR Infodemiology. Mar 24, 2025;5:e67119. [doi: 10.2196/67119] [Medline: 40127460]
23. Fricker M. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press; 2007. [doi: 10.1093/acprof:oso/9780198237907.001.0001] ISBN: 9780198237907
24. Omar M, Sorin V, Wieler LH, et al. Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. Lancet Digit Health. Jan 2026;8(1):100949. [doi: 10.1016/j.landig.2025.100949] [Medline: 41672646]
25. Teng D, Tan L, Cao Q, et al. Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students. NPJ Digit Med. Mar 17, 2026;9(1):356. [doi: 10.1038/s41746-026-02547-z] [Medline: 41844809]
26. Modi ND, Menz BD, Awaty AA, et al. Assessing the system-instruction vulnerabilities of large language models to malicious conversion into health disinformation chatbots. Ann Intern Med. Aug 2025;178(8):1172-1180. [doi: 10.7326/ANNALS-24-03933] [Medline: 40550134]
27. Liu T, Wang P, Pan D, Liu R. Credibility of AI generated and human video doctors and the relationship to social media use. Front Public Health. 2025;13:1559378. [doi: 10.3389/fpubh.2025.1559378]
28. Deep dive on supplement scams: how AI drives 'Miracle Cures' and sponsored health-related scams on social media. Bitdefender. URL: https://www.bitdefender.com/en-us/blog/labs/deep-dive-on-supplement-scams-how-ai-drives-miracle-cures-and-sponsored-health-related-scams-on-social-media [Accessed 2026-07-27]
29. Revealed: how academics are being deepfaked on TikTok and Instagram to promote supplements. Full Fact. 2025. URL: https://fullfact.org/health/academics-deepfaked-tiktok-wellness-nest/ [Accessed 2026-06-27]
30. AMA CEO: deepfake doctors are a threat to public health. STAT News. 2026. URL: https://www.statnews.com/2026/02/17/deepfake-doctors-scam-ama/ [Accessed 2026-02-17]
31. Jafar Z, Quick JD, Larson HJ, et al. Social media for public health: reaping the benefits, mitigating the harms. Health Promot Perspect. 2023;13(2):105-112. [doi: 10.34172/hpp.2023.13] [Medline: 37600540]
32. Mayer RE. Multimedia Learning. 2nd ed. Cambridge University Press; 2009. ISBN: 978-0521735353
33. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. Mar 9, 2018;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
34. Phillips W. The oxygen of amplification. Data & Society. 2018. URL: https://datasociety.net/wp-content/uploads/2018/05/2-PART-2_Oxygen_of_Amplification_DS.pdf [Accessed 2026-07-21]

35. Wack M, Ehrett C, Linvill D, Warren P. Generative propaganda: evidence of AI's impact from a state-backed disinformation campaign. PNAS Nexus. Apr 2025;4(4):pgaf083. [doi: 10.1093/pnasnexus/pgaf083] [Medline: 40171239]
36. Chandrasekaran R, Moustakas E, Sadiq MT. Racial and demographic disparities in susceptibility to health misinformation on social media: national survey-based analysis. J Med Internet Res. Nov 6, 2024;26:e55086. [doi: 10.2196/55086] [Medline: 39504121]
37. AI in the age of fake (imagined) content. Stimson. Feb 2026. URL: https://www.stimson.org/2026/ai-in-the-age-of-fake-imagined-content/ [Accessed 2026-07-21]
38. Freidson E. Profession of Medicine: A Study of the Sociology of Applied Knowledge. Dodd, Mead; 1970. URL: https://archive.org/details/professionofmedi0000frei [Accessed 2026-07-27]
39. Starr P. The Social Transformation of American Medicine. Basic Books; 1982. URL: https://archive.org/details/socialtransforma0000star [Accessed 2026-07-27]
40. Vaccari C, Chadwick A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Soc Media Soc. Jan 2020;6(1):2056305120903408. [doi: 10.1177/2056305120903408]
41. Chandra NA, Murtfeldt R, Qiu L, et al. Deepfake-eval-2024: a multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv. Preprint posted online on Mar 4, 2025. [doi: 10.48550/arXiv.2503.02857]
42. Kozyreva A, Lorenz-Spreen P, Herzog SM, et al. Toolbox of individual-level interventions against online misinformation. Nat Hum Behav. Jun 2024;8(6):1044-1052. [doi: 10.1038/s41562-024-01881-0] [Medline: 38740990]

Abbreviations

C2PA: Coalition for Content Provenance and Authenticity
GenAI: generative AI
LLM: large language model
LMIC: lower- and middle-income countries

Edited by Tina Purnat; peer-reviewed by David Scales, Tina Purnat; submitted 30.Mar.2026; final revised version received 09.Jul.2026; accepted 09.Jul.2026; published 20.Aug.2026

Please cite as:
Malkoç M
Multilayered Epistemic Disruption in AI-Driven Health Misinformation: Conceptual Framework and Viewpoint
JMIR Infodemiology 2026;6:e96664
URL: https://infodemiology.jmir.org/2026/1/e96664
doi: 10.2196/96664