AIR Journal of Life Sciences and Medicine, VOL. 2026
Article Unique ID: AIRLSM2026232
DOI: 10.65737/AIRLSM2026232

A COGNITIVE RISK TAXONOMY FOR AI-ASSISTED EMERGENCY ECG INTERPRETATION: MAPPING ALGORITHMIC ERROR TO PHYSICIAN BIAS PATHWAYS

Al-Baraa Dabaliz[1*], Al-Awwab Dabaliz[2], Mosab Hawarey[3]
[1*] Specialty Doctor Emergency Medicine, Emergency Department UHNM (*al-baraa.dabaliz@uhnm.nhs.uk*)
[2] Assistant Professor of Clinical Anesthesiology, University of Illinois Chicago
[3] Director, Geospatial Research

Received: 2026 February 12 | Approved: 2026 February 14 | Published: 2026 February 15

Abstract

Background: Artificial intelligence algorithms for electrocardiogram (ECG) interpretation are now standard in emergency departments worldwide, yet the assumption that AI and physician errors are complementary — and therefore self-correcting — has never been systematically tested against the cognitive realities of emergency medicine practice.
Objective: To develop a cognitive risk taxonomy that maps specific AI error patterns through specific physician bias pathways to specific clinical risk predictions for emergency-critical ECG diagnoses.
Methods: We analysed 21,799 dual-labelled ECGs from the PTB-XL+ dataset, comparing expert cardiologist annotations against GE 12SL algorithm output across 54 emergency-critical SNOMED CT concepts spanning acute myocardial infarction, life-threatening arrhythmias, conduction blocks, and ischaemia. A five-stage disagreement analysis framework quantified error direction, magnitude, confidence profiles, compound co-occurrence patterns, and diagnostic substitution profiles. Each disagreement signature was mapped to cognitive bias pathways derived from dual-process theory and scored on a composite of frequency, clinical severity, and bias amplification potential.
Results: Of 106,401 non-trivial comparisons, 93.7% were discordant (6.3% agreement), with 81% of the dataset carrying at least one emergency-critical disagreement (mean 5.7 per affected ECG). Eight named disagreement signatures were identified, organised into a two-tier, four-class taxonomy: Lethal Diagnosis Miss (anteroseptal MI blind spot, SVT/VT inversion, LBBB–STEMI mask), Mechanism Blindness (bradycardia inflation, fascicular desert), Signal Corruption (ischaemia noise floor, old MI avalanche), and Self-Undermining AI (QT alarm fatigue). Seven were classified as CRITICAL risk. The AI's binary confidence architecture delivered 81% of overcalls at maximum confidence with no hedging, creating near-maximum anchoring potential for every error. Diagnostic substitution profiling revealed that AI misses are not silent omissions but active reframings: when the AI misses anteroseptal MI, it labels "Old MI" on 60.6% of those ECGs; when it misses ventricular tachycardia, it labels "SVT" on 45.8%.
Conclusions: The Compound Risk Hypothesis was validated across all eight danger zones: in every case, the AI–physician combination was projected to perform worse than the physician alone through three mechanisms — Direct Suppression, Capability Destruction, and

Environmental Contamination. A seven-component safeguard framework targeting class-specific bias pathways was developed, designed to be net-negative on alert burden. The taxonomy framework is system-agnostic and designed for reapplication to any AI system with dual-label validation data.

Keywords: electrocardiogram; artificial intelligence; cognitive bias; automation bias; clinical decision support; emergency medicine; diagnostic error; anchoring; human–AI interaction; patient safety.

Citation: Dabaliz, Al-Baraa, Dabaliz, Al-Awwab, Hawarey, Mosab (2026), A Cognitive Risk Taxonomy for AI-Assisted Emergency ECG Interpretation: Mapping Algorithmic Error to Physician Bias Pathways, AIR Journal of Life Sciences and Medicine, Vol. 2026, AIRLSM2026232, DOI: 10.65737/AIRLSM2026232.

1. Introduction

Electrocardiogram interpretation is among the most time-critical diagnostic tasks in emergency medicine. In patients presenting with acute coronary syndromes, the interval between ECG acquisition and clinical decision directly determines myocardial salvage and survival, with guidelines mandating door-to-ECG times of ten minutes or less and door-to-balloon times of ninety minutes or less (O'Gara et al., 2013). To support this high-stakes workflow, commercial AI-based ECG interpretation algorithms have been deployed across millions of ECG machines worldwide. The GE 12SL algorithm alone is installed on the majority of hospital ECG systems globally, providing automated diagnostic statements that emergency physicians encounter before or simultaneously with their own reading of the tracing. The prevailing assumption underlying this deployment is one of complementary error: that AI and human interpreters make independent, non-overlapping mistakes, such that combining both perspectives yields diagnostic performance superior to either alone. This assumption has not been rigorously tested against the cognitive science of physician decision-making under emergency department conditions.

A substantial body of evidence from cognitive psychology and emergency medicine demonstrates that the emergency department creates a uniquely bias-prone cognitive environment. Dual-process theory, the dominant framework for understanding clinical cognition, distinguishes between fast, intuitive pattern recognition (System 1) and slow, analytical deliberation (System 2) (Croskerry, 2009). Emergency departments maximally load the conditions that degrade System 2 processing—time pressure, cognitive load, information incompleteness, and consequence asymmetry—while simultaneously demanding the highest-stakes decisions. Within this environment, anchoring bias has been documented in 82% of triage decisions (Rana et al., 2025), automation bias causes clinicians to override their own correct judgments in favour of erroneous computer recommendations in 6% of cases (Goddard et al., 2012), and premature closure has been identified as the single most common cause of missed diagnoses (Graber, 2013). When an AI system introduces a confident diagnostic interpretation into this cognitively vulnerable environment, it does not function as a neutral second opinion. It functions as a powerful System 1 anchor—an immediate frame that shapes all subsequent reasoning about the ECG. If the AI is wrong, this anchor activates a predictable cascade of

anchoring, confirmation bias, premature closure, and automation bias that can drive the physician toward the same error the AI made.

Prior work has characterised the general phenomenon of automation bias in ECG interpretation (Bond et al., 2018), the accuracy limitations of commercial algorithms (Schläpfer and Wellens, 2017), and the susceptibility of clinicians to AI-driven decision-support errors across medical domains (Gaube et al., 2021; Parasuraman and Manzey, 2010). However, no study has systematically mapped *specific* AI error patterns on *specific* emergency-critical diagnoses to *specific* cognitive bias pathways in order to predict where AI assistance degrades rather than improves physician performance. This gap is consequential because the danger of AI error lies not in the error itself but in the cognitive bias pathway it activates in the receiving physician. Furthermore, recent experimental evidence demonstrates that humans do not merely defer to biased AI in the moment; they internalise AI biases, reproducing the AI's systematic errors even after the AI is removed (Vicente and Matute, 2023). This transforms what might appear to be a transient technological problem into a durable cognitive contamination—one that persists beyond the AI-assisted encounter and potentially degrades unassisted diagnostic performance in future patients.

In this study, we develop a cognitive risk taxonomy for AI–human disagreement in emergency ECG interpretation. Using the PTB-XL+ dataset's dual-label architecture (21,799 ECGs with paired commercial AI and expert cardiologist annotations), we analyse disagreement patterns across 54 emergency-critical SNOMED-coded diagnoses organised into nine clinical categories. We identify eight named disagreement signatures, map each to cognitive bias pathways derived from dual-process theory and the automation bias literature, score them for compound clinical risk using a weighted composite metric, and organise them into a hierarchical taxonomy of four danger classes. We test the Compound Risk Hypothesis: that in identifiable danger zones, the AI–physician combination is predicted to perform worse than the physician alone, because the AI's errors systematically activate rather than counteract physician cognitive biases. We propose a class-specific safeguard architecture designed to interrupt the identified bias pathways. Although illustrated here using the GE 12SL algorithm as a widely deployed exemplar, the analytical framework—the mapping from AI error profiles through cognitive bias pathways to compound clinical risk—is designed to generalise to any AI ECG interpretation system whose error characteristics can be empirically profiled.

2. Methods

2.1 Data Source and Study Population

This study uses the PTB-XL+ dataset (Strodthoff et al., 2023), a publicly available extension of the PTB-XL electrocardiogram corpus (Wagner et al., 2020). PTB-XL+ provides 21,799 ten-second, 12-lead ECGs, each carrying two independent sets of diagnostic labels mapped to a common SNOMED CT ontology: expert cardiologist annotations serving as the reference standard (derived from the original PTB-XL clinical reports) and automated interpretations generated by the GE 12SL version 23 commercial algorithm serving as the index test. All 21,799 ECGs have both label sets with complete SNOMED mapping and zero missing records. The SNOMED ontology encompasses 287 total diagnostic concepts, of which 162 are informative

(assigned at least once by either source) and 93 are present in both the AI and human label sets with sufficient frequency for meaningful comparison.

The GE 12SL is a rule-based algorithm originating in the mid-2000s and remains the most widely deployed commercial ECG interpretation system globally. We selected it for analysis because it represents the AI output that the majority of emergency physicians currently encounter in clinical practice, and because PTB-XL+ provides the only publicly available dual-label dataset with matched SNOMED coding from both a commercial algorithm and expert cardiologists. We recognise that the 12SL is a legacy system whose error profile may not represent contemporary deep neural network approaches; this limitation is addressed in the Discussion, and the analytical framework developed here is designed for reapplication to any AI system whose error characteristics can be empirically profiled.

2.2 Emergency-Critical Concept Selection

From the 93 informative SNOMED concepts present in both label sets, we selected 54 emergency-critical concepts using five clinical inclusion criteria: time-critical diagnoses where delayed recognition directly increases mortality (acute myocardial infarction subtypes, ST-elevation, life-threatening arrhythmias); diagnoses where AI-human disagreement creates specific cognitive risk in emergency settings (old myocardial infarction overcalls that anchor physicians, STEMI mimics); conduction abnormalities with risk of deterioration to complete heart block or sudden death; diagnoses on the acute coronary syndrome spectrum where emergency department disposition decisions hinge on ECG interpretation; and low-frequency but catastrophic-if-missed conditions including pre-excitation and QT prolongation. The 54 selected concepts were organised into nine clinical categories: Acute MI (8 concepts), Old MI (3), MI Localisation (8), Ischemia/ACS Spectrum (10), Life-Threatening Arrhythmias (10), Conduction Blocks (11), Hypertrophy (2), Pre-excitation (1), and QT Prolongation (1). The complete list of all 93 informative concepts with inclusion and exclusion rationale is provided in Supplementary Table S4.

2.3 Disagreement Analysis Framework

The disagreement analysis proceeded in five stages, each producing data structures that feed into the subsequent taxonomy construction.

First, record-level binary comparison was performed for each combination of ECG and SNOMED concept across all 21,799 ECGs and 54 concepts, yielding 106,401 non-trivial comparison records (those where at least one source assigned the concept as positive). Each record was classified as agreement (both sources positive), AI overcall (AI positive, human negative), or AI miss (AI negative, human positive).

Second, inter-rater agreement statistics were computed for all 54 concepts. Cohen’s kappa was calculated using direct source labels (the SNOMED concepts explicitly assigned by each source without ontological hierarchy propagation), as this measures whether the two raters make the same diagnostic claim rather than whether their claims are ontologically related. McNemar’s test with Yates continuity correction assessed whether the directional asymmetry between overcalls

and misses was statistically significant for each concept. The direct-label approach provides a conservative estimate of disagreement; the hierarchy-inclusive approach used in the clinical impact analyses captures the full scope of diagnostic influence.

Third, confidence stratification was performed by cross-tabulating AI confidence scores (reported on a 0 to 100% scale) against disagreement direction for each concept. This identified the proportion of AI errors delivered at maximum confidence, which determines the strength of the anchoring signal received by the interpreting physician.

Fourth, compound co-occurrence analysis examined disagreement patterns at the individual ECG level. For each ECG carrying at least one disagreement, we identified whether it simultaneously carried overcalls and misses across different diagnostic categories. Six named compound danger patterns were characterised in which the AI overcalls a lower-acuity finding while simultaneously missing a higher-acuity one on the same tracing.

Fifth, diagnostic substitution profiling determined what the AI labels instead when it misses a specific diagnosis. For each ECG with an AI miss on a given concept, we identified which other emergency-critical concepts the AI assigned as positive on that same ECG. This reveals whether AI misses are silent omissions or active competing narratives that provide an alternative (incorrect) diagnostic frame.

Additionally, SNOMED hierarchy decomposition was performed to separate leaf-level concepts (the most specific diagnostic statements) from parent-level aggregate concepts, controlling for ontological inflation in which a single clinical finding generates concordant records at multiple levels of the SNOMED hierarchy.

2.4 Cognitive Risk Taxonomy Construction

The taxonomy was constructed through a four-step pipeline that maps empirical disagreement data to cognitive risk predictions.

In the first step, signature identification, concepts with similar error profiles were clustered into named disagreement signatures based on four dimensions: dominant error direction (overcall versus miss), error magnitude (AI-to-human label count ratio), confidence profile (proportion of errors at maximum AI confidence), and compound pattern involvement (co-occurrence with errors in other diagnostic categories). Eight signatures emerged from this synthesis, each representing a distinct pattern of AI failure with specific cognitive implications.

In the second step, error-to-bias mapping, each signature was mapped to cognitive bias pathways derived from dual-process theory (Croskerry, 2009) and the automation bias literature (Goddard et al., 2012; Parasuraman and Manzey, 2010). Four pathways were defined: Pathway A, in which AI overcalls anchor the physician toward a false-positive diagnosis through anchoring, confirmation bias, and premature closure; Pathway B, in which AI misses suppress physician detection through automation omission bias and false-reassurance anchoring; Pathway C, in which high-confidence AI errors maximise anchor strength and suppress System 2 analytical engagement; and Pathway D, in which mass overcalling shifts the physician's calibration for

disease prevalence across patients. For each signature, the mapping involved three steps: classification of the dominant error type, determination of which pathways the error activates given the confidence profile, and construction of the specific ordered sequence of cognitive biases activated within each pathway.

In the third step, composite risk scoring, each signature was scored on three weighted dimensions: error frequency (F, weight 0.30), reflecting the volume of affected ECGs on a log-transformed scale; clinical severity (S, weight 0.40), reflecting the time-criticality and mortality impact of the affected diagnoses; and bias amplification potential (B, weight 0.30), reflecting the strength and number of cognitive biases activated, weighted by confidence level and compound pattern presence. The severity dimension received the highest weight on the principle that a rare error on a lethal diagnosis ranks higher than a frequent error on a non-lethal one. Composite scores yielded two tiers: CRITICAL (score 8.0 or above) and HIGH (score 6.5 to 7.9).

In the fourth step, the Compound Risk Hypothesis was tested for each signature by comparing the projected performance of the AI-assisted physician against the unassisted physician baseline, using literature-derived parameters for physician detection rates and automation bias effect sizes. The hypothesis predicts that in identifiable danger zones, the AI–physician combination performs worse than the physician alone because the AI's errors systematically activate rather than counteract physician cognitive biases.

3. Results

3.1 Overall Disagreement Landscape

Across 21,799 ECGs and 54 emergency-critical SNOMED concepts, 106,401 non-trivial comparison records were generated (those where at least one source assigned the concept as positive). Of these, 99,734 (93.7%) were discordant: 52,544 AI overcalls and 47,190 AI misses, with only 6,667 records (6.3%) showing positive agreement between AI and expert cardiologists. When analysis was restricted to the 41 leaf-level concepts (the most specific, clinically actionable diagnostic statements), concordance dropped further to 3.0%, confirming that disagreement intensifies at the level where clinical decisions are actually made.

Cohen's kappa across the 54 concepts yielded a mean of 0.314 (fair agreement) and a median of 0.283, with a range from −0.015 (worse than chance) to 0.890. Only four concepts achieved almost perfect agreement ($\kappa > 0.80$), while 21 showed only slight agreement or worse ($\kappa < 0.20$). McNemar's test demonstrated statistically significant directional asymmetry ($p < 0.001$) for 45 of 54 concepts (83%), confirming that the disagreement is not symmetric noise but systematic directional error. Table 1 summarises these findings.

The disagreement was not isolated to individual ECG-concept pairs but clustered heavily at the ECG level. A total of 17,650 ECGs (81.0% of the dataset) carried at least one emergency-critical disagreement. Among these affected ECGs, the mean number of simultaneous disagreements was 5.7 (median 5, maximum 20), and 78% carried three or more concurrent errors. This density of per-ECG error is clinically consequential: the emergency physician confronting an AI-

interpreted ECG is typically facing not one potential AI error but a cascade of five to six, each activating its own cognitive bias pathway under time pressure.

Table 1. Summary of AI–human disagreement across 54 emergency-critical SNOMED concepts. Agreement and disagreement rates are computed from 106,401 non-trivial comparison records (hierarchy-inclusive) and 50,039 records (leaf-only). Kappa values are based on direct source labels across all 21,799 ECGs. McNemar's test assesses directional asymmetry between AI overcalls and AI misses per concept.

Metric	Value
Total non-trivial comparisons	106,401
AI overcalls	52,544 (49.4%)
AI misses	47,190 (44.3%)
Positive agreements	6,667 (6.3%)
Leaf-only disagreement rate	97.0%
Mean Cohen's κ (54 concepts)	0.314 (fair)
Concepts with $\kappa < 0.20$	21 (38.9%)
Concepts with significant McNemar's ($p < 0.001$)	45 (83.3%)
ECGs with ≥1 ER-critical disagreement	17,650 (81.0%)
Mean disagreements per affected ECG	5.7

3.2 Directional Asymmetry and Confidence in Error

The 54 concepts divided into 29 dominated by AI overcall, 22 dominated by AI miss, and 3 with balanced bidirectional error. The magnitude of directional asymmetry was extreme at the tails. The highest-magnitude overcalls were old myocardial infarction (AI-to-human label count ratio 47.6×), supraventricular tachycardia (16.2×), old inferior MI (41.8×), tachyarrhythmia (10.6×), and lateral ischemia on ECG (9.2×). The highest-magnitude misses were acute inferolateral MI (0.04×, representing near-complete AI blindness), anteroseptal infarction on ECG (0.2×), left posterior fascicular block (0.1×), ventricular tachycardia (0.3×), and left anterior fascicular block (0.3×). Figure 1 displays this landscape as a log-scale scatter plot, with extreme outliers labelled.

The confidence profile of AI errors was the single most consequential finding for the cognitive risk analysis. Of all 52,544 AI overcalls, 42,554 (81.0%) were delivered at maximum confidence (100%). The distribution was essentially bimodal: 81% at maximum confidence, zero in the 75–99% range, and 19% distributed across lower confidence bands. By diagnostic category, conduction block overcalls were at 100% maximum confidence in every instance, arrhythmia overcalls at 99.8%, and acute MI overcalls at 91.5%. This binary confidence profile means that the AI never signals uncertainty when it is wrong. Per the cognitive framework established in the Introduction, maximum-confidence errors create the strongest possible anchoring signal and maximally suppress the physician's System 2 analytical engagement, producing the conditions under which automation bias is most potent.

Conversely, when the AI missed findings that expert cardiologists identified, the humans did so with high confidence: 76.0% of all AI misses corresponded to findings identified by cardiologists at 100% confidence, rising to 97–98% for conduction blocks. This establishes that the AI is

systematically failing to detect findings that expert cardiologists identify with near-certainty, not borderline cases where reasonable disagreement might be expected.

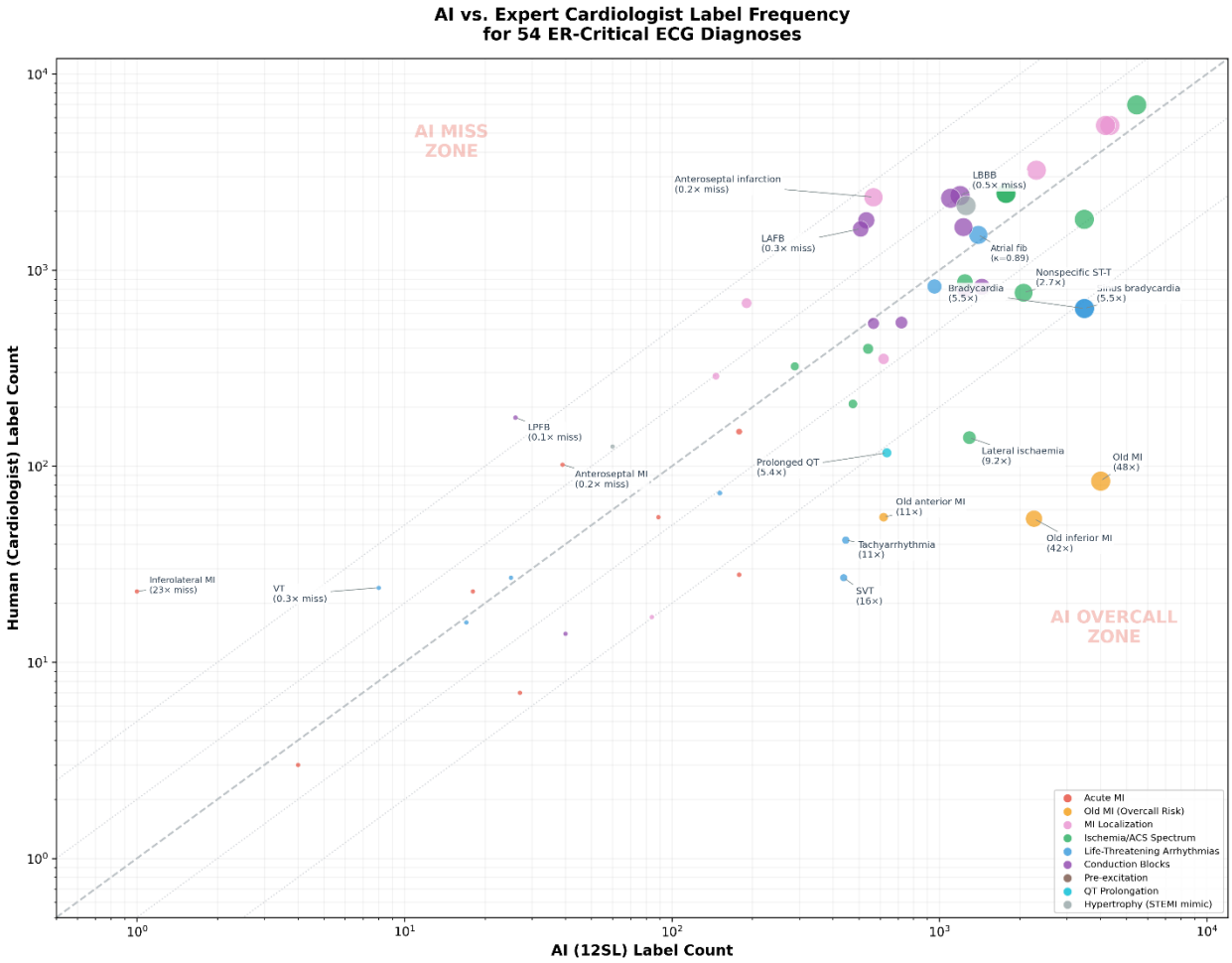

Figure 1. Overcall–miss asymmetry across 54 emergency-critical SNOMED concepts. Each point represents one concept, plotted on log-scaled axes with AI label count on the horizontal axis and human expert label count on the vertical axis. The diagonal line indicates equal labelling frequency. Points below the diagonal represent AI overcall; points above represent AI miss. Colour indicates clinical category. Extreme outliers are labelled: Old MI (47.6× overcall), SVT (16.2× overcall), Anteroseptal MI on ECG (0.2× detection), and LPFB (0.1× detection).

3.3 Eight Disagreement Signatures

Synthesis of the directional, confidence, compound, and hierarchy analyses yielded eight named disagreement signatures, each representing a distinct pattern of AI failure with specific cognitive risk implications.

SIG-1, the Old MI Avalanche, is a mass overcall pattern in which the AI labels old myocardial infarction 47.6 times more frequently than expert cardiologists, generating approximately 6,800 false positives at 72% maximum confidence. In an emergency department chest pain evaluation, this anchor shifts the diagnostic frame from “acute event” to “old problem,” potentially delaying recognition of evolving infarction. On 768 ECGs, this overcall co-occurs with a miss of MI

localisation, meaning the AI labels “old MI” while failing to identify where the infarction actually is.

SIG-2, the Anteroseptal Blind Spot, is a systematic miss in which the AI detects anteroseptal infarction at only 0.2 times the expert rate, missing approximately 2,400 findings that cardiologists identify at 98% confidence. The anteroseptal territory corresponds to left anterior descending artery distribution, the highest-mortality coronary anatomy.

SIG-3, the Bradycardia Inflation, is a mass overcall of heart rate (5.5×) with simultaneous blindness to the underlying conduction mechanism. On 621 ECGs, the AI overcalls bradycardia while missing a conduction block on the same tracing, labelling the rate but missing the pathology.

SIG-4, the SVT/VT Inversion, is a directional swap in which the AI overcalls supraventricular tachycardia 16.2 times while detecting ventricular tachycardia at only 0.3 times the expert rate, systematically labelling a lethal arrhythmia as a benign one. All SVT overcalls are at 100% confidence.

SIG-5, the Ischemia Noise Floor, is a signal dilution pattern in which AI generates approximately 6,500 false ischemia calls (lateral ischemia overcalled 9.2×, nonspecific ST-T changes 2.7×), creating a cry-wolf environment that desensitises physicians to genuine ischaemic findings. On 321 ECGs, a nonspecific overcall co-occurs with a specific ischemia miss.

SIG-6, the LBBB–STEMI Mask, is a systematic miss in which the AI detects left bundle branch block at only 0.5 times the expert rate, missing approximately 2,300 findings. New LBBB with chest pain constitutes a STEMI-equivalent requiring catheterisation laboratory activation; AI blindness to LBBB directly removes the trigger for this critical clinical pathway.

SIG-7, the Fascicular Desert, represents near-total AI blindness to fascicular blocks (left anterior fascicular block 0.3×, left posterior fascicular block 0.1×), missing approximately 3,500 findings. New fascicular block in an acute presentation signals proximal conduction disease with risk of progression to complete heart block.

SIG-8, QT Alarm Fatigue, is an overcall pattern (5.4×, 100% confidence) affecting a measurement where physicians have low baseline sensitivity (36–62%). The overcalling destroys through alarm fatigue the trust needed for the AI’s potential contribution in a domain where it could genuinely add value.

3.4 Compound Patterns and Diagnostic Substitution

Compound analysis at the ECG level revealed that 6,661 affected ECGs (37.7%) carried simultaneous AI overcalls and AI misses on the same tracing. This bidirectional error pattern means that on more than a third of disagreement-bearing ECGs, the AI simultaneously anchors the physician on a finding that is not present while failing to flag a finding that is. Six named compound danger patterns were identified, the most frequent being old MI overcall co-occurring with MI localisation miss (768 ECGs) and bradycardia overcall co-occurring with conduction

block miss (621 ECGs). Cross-category analysis showed that arrhythmia overcalls masking ischemia misses was the single most common inter-category compound pattern, affecting 1,684 ECGs.

Diagnostic substitution profiling revealed that AI misses are not silent omissions but active reframings. Table 2 presents the substitution profiles for the four most clinically dangerous miss signatures. When the AI misses anteroseptal MI, it labels "Old MI" on 60.6% of those ECGs, reframing an acute territory-specific infarction pattern as chronic disease. When it misses ventricular tachycardia, it labels "SVT" or "Tachyarrhythmia" on 45.8% and 54.2% of those ECGs respectively, directly confirming the SVT/VT inversion. When it misses LBBB, it labels "Old MI" on 32.4%, interpreting LBBB-associated morphological changes as prior infarction rather than conduction abnormality. These substitution profiles transform errors of omission into errors of commission: the AI does not merely fail to detect a dangerous finding but provides a competing diagnostic narrative that anchors the physician on a less urgent or incorrect interpretation.

Table 2. Diagnostic substitution profiles for key AI miss signatures. For each signature, the table shows what the AI labels instead on ECGs where it misses the target diagnosis. Percentages indicate the proportion of AI-miss ECGs on which the substitute label appears. Multiple substitutes may co-occur on the same ECG.

Signature	Missed Diagnosis (n ECGs)	Primary AI Substitute	% of Miss ECGs	Clinical Reframing
SIG-2	Anteroseptal MI on ECG (1,814)	Old MI	60.6%	Acute → chronic
SIG-4	Ventricular tachycardia (24)	SVT / Tachyarrhythmia	45.8% / 54.2%	Lethal → benign rhythm
SIG-6	Left bundle branch block (1,403)	Old MI	32.4%	Conduction → infarction
SIG-7	Left posterior fascicular block (166)	RBBB / Complete RBBB	65.7% / 57.8%	Half the bifascicular pattern

4. The Cognitive Risk Taxonomy

4.1 Error-to-Bias Mapping

The eight disagreement signatures identified in Section 3 were each mapped to cognitive bias pathways derived from dual-process theory. Four pathways were defined based on the interaction between AI error type and physician cognitive processing. Pathway A is activated by AI overcalls: the false-positive diagnosis anchors the physician, triggers selective attention to confirming features via confirmation bias, and terminates the diagnostic search through premature closure. Pathway B is activated by AI misses: the absence of an AI flag suppresses physician detection through automation omission bias, with the implicit "normal" frame reinforcing premature closure on a benign interpretation. Pathway C is activated by high-confidence errors regardless of direction: when the AI delivers a wrong diagnosis at maximum confidence, it creates the strongest possible anchoring signal and maximally suppresses System 2 analytical engagement. Pathway D operates across patients rather than within a single ECG:

mass overcalling shifts the physician's mental model of disease prevalence, degrading calibration for future encounters through the availability heuristic.

All eight signatures activated Pathway C, a consequence of the 12SL algorithm's binary confidence architecture (81% of overcalls at 100% confidence, zero in the 75–99% range). This means that every AI error in the dataset arrives with near-maximum anchoring power, leaving the physician no confidence gradient on which to calibrate trust.

The SVT/VT Inversion (SIG-4) illustrates the mapping at its most dangerous. This signature simultaneously activates Pathway A (the 16.2× SVT overcall anchors the physician toward a benign arrhythmia) and Pathway B (the 0.3× VT miss suppresses detection of the lethal rhythm) on the same ECG. The physician receives a maximally confident label for a tolerable arrhythmia while the life-threatening diagnosis is absent from the AI output. The resulting cognitive state combines compound anchoring toward the benign entity with compound omission of the lethal one, producing what we term Bidirectional Pathway Activation. No prior theoretical model of AI–human cognitive interaction describes this simultaneous dual-pathway mechanism from a single diagnostic output.

Three additional novel interaction patterns emerged from the mapping that were not anticipated in the theoretical framework. Clinical Pathway Interruption (SIG-6) occurs when the AI miss removes the entry condition for an evidence-based protocol: AI blindness to LBBB does not merely omit a diagnosis but eliminates the trigger for catheterisation laboratory activation in a STEMI-equivalent presentation, extending the cognitive consequence from diagnostic omission to protocol-level failure. The Delegation Trap (SIG-7) describes a paradox in which physician delegation is highest precisely where AI capability is lowest: axis determination is perceived as a computational task ideally suited to automation, increasing physician reliance, yet the AI shows near-total blindness to the pathological axis deviations that define fascicular blocks. The Self-Undermining Loop (SIG-8) occurs when the AI overcalls a measurement it is objectively better positioned to perform: physician baseline sensitivity for QT prolongation is only 36–62% (Viskin et al., 2005), making the AI a potential net contributor, but its 5.4× overcall rate generates alarm fatigue that destroys the trust required for its own contribution. The complete error-to-bias mapping for all eight signatures is provided in Supplementary Table S3.

4.2 Composite Risk Scoring and Classification

Each signature was scored on three weighted dimensions: error frequency (F, weight 0.30), reflecting the volume of affected ECGs on a log-transformed scale; clinical severity (S, weight 0.40), reflecting the time-criticality and mortality impact of the affected diagnoses; and bias amplification potential (B, weight 0.30), reflecting the number and strength of cognitive biases activated, weighted by confidence level and compound pattern presence. Composite scores ranged from 7.1 to 9.1. Seven signatures were classified as CRITICAL (composite ≥ 8.0) and one as HIGH (composite 6.5–7.9). Table 3 presents the complete taxonomy.

The bias amplification dimension showed limited differentiation across signatures (range 8.2–9.8 after refinement for pathway complexity), a finding that reflects a genuine property of the 12SL system rather than a limitation of the scoring method. The algorithm's binary confidence profile,

in which 81% of all overcalls are delivered at maximum confidence with no hedging, creates a floor condition where every error activates near-maximum anchoring potential. A better-calibrated AI system would be expected to show greater B-score variation across signatures, potentially altering the rank ordering.

Table 3. The Cognitive Risk Taxonomy for AI-assisted emergency ECG interpretation. Eight danger zones are organised into two tiers and four classes by primary risk mechanism. Composite risk score = 0.30 × Frequency + 0.40 × Severity + 0.30 × Bias Amplification. CRITICAL tier: composite ≥ 8.0; HIGH tier: 6.5–7.9. Pathway indicates the cognitive bias pathway(s) activated. Affected ECGs includes compound pattern ECGs where applicable.

Tier	Class	#	Danger Zone	Score	Path	Bias Chain	ECGs	Worst-Case Outcome
CRIT	A: Lethal Dx Miss	1	**SIG-2: Anteroseptal Blind Spot**	9.1	B+C	Omission → Normal anchor → Closure	~2,400	Missed widowmaker MI → death
CRIT	A	2	**SIG-4: SVT/VT Inversion**	8.7	A+B+C	Dual anchor (benign) + Dual omission (lethal)	~462	VT as SVT → VF → arrest
CRIT	A	3	**SIG-6: LBBB–STEMI Mask**	8.9	B+C	Omission → Protocol interrupt → No cath lab	~2,300	Missed STEMI-equiv → death
CRIT	B: Mechanism Blindness	4	**SIG-3: Bradycardia Inflation**	9.1	A+C	Rate anchor → Mechanism closure	~4,021	Block untreated → complete HB
CRIT	B	5	**SIG-7: Fascicular Desert**	8.7	B+C	Delegation trap → Axis anchor → Omission	~3,500	Missed progression → sudden block
CRIT	C: Signal Corruption	6	**SIG-5: Ischemia Noise Floor**	8.4	A→D	Cry wolf → Desensitisation → Future miss	~6,821	Real ischaemia dismissed → MI
CRIT	C	7	**SIG-1: Old MI Avalanche**	8.3	A+C	Max anchor → “Old problem” frame	~7,568	Acute MI reframed as chronic
HIGH	D: Self-Undermining AI	8	**SIG-8: QT Alarm Fatigue**	7.1	A→D	Alarm fatigue → Aversion → Lost QT	~631	Torsades when AI dismissed

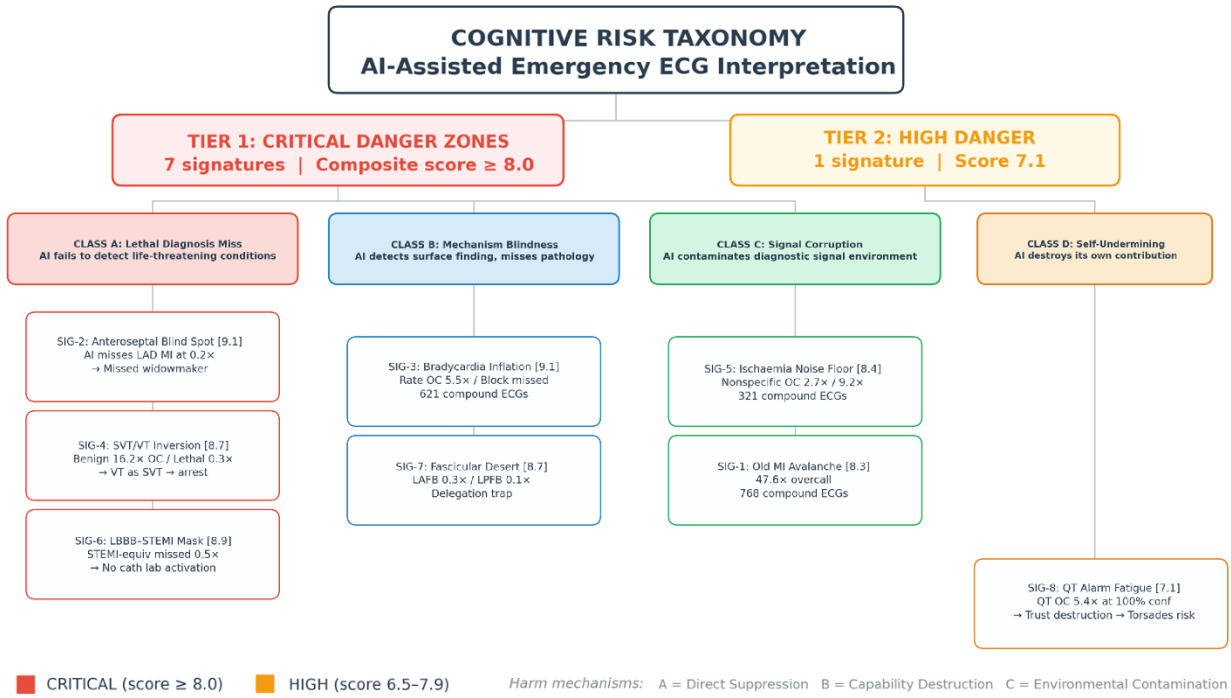

Figure 2. Hierarchical structure of the Cognitive Risk Taxonomy. The taxonomy is organised as a tree with two tiers (CRITICAL and HIGH), four classes (A: Lethal Diagnosis Miss, B: Mechanism Blindness, C: Signal Corruption, D: Self-Undermining AI), and eight named danger zones. Composite risk scores are shown in brackets.

4.3 Taxonomy Structure

The four classes capture distinct cognitive risk mechanisms that require different safeguard strategies.

Class A, Lethal Diagnosis Miss, comprises three signatures (SIG-2, SIG-4, SIG-6) in which the AI fails to detect immediately life-threatening conditions. The unifying mechanism is Pathway B activation: the AI's silence on a dangerous diagnosis suppresses physician detection through automation omission bias. In the emergency department time-pressure environment, where System 2 analytical capacity is already depleted, these omissions are rarely recovered through independent physician assessment. The class includes the highest-risk signature in the taxonomy (SIG-2, Anteroseptal Blind Spot, composite 9.1), reflecting the convergence of high clinical severity (left anterior descending artery territory MI), extreme AI miss magnitude (0.2× detection rate), and strong automation omission bias activation.

Class B, Mechanism Blindness, comprises two signatures (SIG-3, SIG-7) in which the AI provides a superficially correct observation that satisfies the physician's initial query while concealing the clinically important underlying mechanism. The AI labels the heart rate but misses the conduction block; it provides an axis calculation but misses the fascicular block that defines the pathological deviation. The surface finding creates premature closure before the deeper mechanistic analysis occurs.

Class C, Signal Corruption, comprises two signatures (SIG-1, SIG-5) in which mass overcalling contaminates the physician's diagnostic calibration across patients. Unlike Classes A and B, where harm is immediate and ECG-specific, Class C harm is cumulative and temporal: it degrades the physician's calibration for disease prevalence over time through the availability heuristic and desensitisation. The 47.6× old MI overcall rate, if internalised through repeated exposure, would create an inflated mental model of MI prevalence that persists even after the AI is removed, consistent with the bias inheritance mechanism described by Vicente and Matute (2023).

Class D, Self-Undermining AI, contains one signature (SIG-8) representing the unique case where the AI overcalls a finding it is objectively better positioned to measure than the physician, destroying through miscalibration the trust required for its own contribution.

Three distinct harm mechanisms operate across the taxonomy: Direct Suppression (Class A), in which AI error reduces physician detection below the unassisted baseline through automation bias; Capability Destruction (Class B and D), in which AI failure at a delegated task removes the physician's independent capability in that domain; and Environmental Contamination (Class C), in which AI error degrades the diagnostic environment across patients through calibration distortion.

4.4 Validation of the Compound Risk Hypothesis

The Compound Risk Hypothesis, proposed in the Introduction, predicts that in identifiable danger zones, the AI–physician combination performs worse than the physician alone because AI errors systematically activate rather than counteract physician cognitive biases. For each signature, we compared the projected AI-assisted physician performance against the unassisted physician baseline using literature-derived parameters.

For Class A signatures, the comparison is direct. Emergency physician sensitivity for subtle STEMI is approximately 65% (Smith et al., 2016). An ideal AI would increase this toward 90% or above by catching subtle patterns the physician misses. However, when the AI misses anteroseptal MI at 0.2× the expert detection rate and delivers "Normal" output, automation omission bias (Goddard et al., 2012) suppresses the physician's independent detection, reducing net sensitivity below the unassisted 65% baseline. For SIG-4, physician SVT/VT discrimination accuracy of approximately 70–80% is converted by the AI's systematic labelling of VT as SVT into a unidirectional error toward the benign diagnosis, reducing discrimination below the unassisted baseline. For SIG-6, AI blindness to LBBB at 0.5× the expert rate removes the Sgarbossa criteria trigger for STEMI-equivalent catheterisation laboratory activation in patients where the physician would otherwise have detected LBBB independently.

In all eight danger zones, the projected net effect of AI assistance was negative: the physician–AI combination was predicted to perform worse than the physician alone. In seven, the harm was unambiguous; in SIG-8 (QT Alarm Fatigue), the effect was mixed but net negative, as the AI's potential benefit in a domain of poor physician baseline performance was self-undermined by overcalling. The complementary error assumption—that AI and physician errors are independent

and cancel—fails in these danger zones because the AI's errors are directionally aligned with the physician's cognitive vulnerabilities, creating reinforcing rather than correcting error.

5. Discussion

5.1 Principal Findings and Significance

This study presents the first cognitive risk taxonomy for AI–human disagreement in emergency ECG interpretation, mapping specific AI error patterns through specific cognitive bias pathways to specific clinical risk predictions. Three findings carry particular significance for the field.

First, the taxonomy itself provides a structured framework for characterising how AI errors interact with physician cognition. Prior work has established the general phenomenon of automation bias in ECG interpretation (Bond et al., 2018) and the accuracy limitations of commercial algorithms (Schläpfer and Wellens, 2017), but these have remained separate literatures. The taxonomy bridges them by demonstrating that the clinical danger of an AI error depends not on the error's magnitude alone but on the cognitive bias pathway it activates in the receiving physician. The same AI miss that might be safely caught by an unhurried cardiologist reviewing a tracing at leisure becomes clinically dangerous when it suppresses detection through automation omission bias in a time-pressured emergency physician operating at the limits of System 1 processing.

Second, diagnostic substitution reframes AI misses as active competing narratives rather than passive omissions. When the AI misses anteroseptal MI, it does not merely output silence; on 60.6% of those ECGs it labels "Old MI," providing a confident alternative diagnostic frame that anchors the physician on chronic disease rather than an evolving acute event. When it misses ventricular tachycardia, it labels "SVT" on 45.8% of those tracings, actively directing the physician toward the benign arrhythmia. This transforms what would be a correctable error of omission into a far more resistant error of commission, because the physician must not only independently detect the missed finding but actively override a confident AI-provided alternative explanation.

Third, validation of the Compound Risk Hypothesis across all eight danger zones challenges the prevailing complementary error assumption that underpins current AI-assisted diagnostic workflows. In every identified danger zone, the AI–physician combination was projected to perform worse than the physician alone. The mechanisms of this harm were not uniform but fell into three categories — Direct Suppression (AI error reduces detection below unassisted baseline), Capability Destruction (AI failure at a delegated task removes the physician's independent capability), and Environmental Contamination (AI error degrades calibration across patients over time). The bias inheritance evidence from Vicente and Matute (2023) suggests that the third mechanism, Environmental Contamination, may persist even after the AI system is removed, transforming a technological problem into a durable cognitive one.

5.2 Safeguard Implications

The taxonomy's class structure directly informs a four-layer safeguard architecture targeting

different points in the cognitive pathway: output restructuring (modifying what the physician sees before reasoning begins), forced analytical engagement (interrupting System 1 acceptance of AI output), omission detection (catching what AI-influenced reasoning missed), and environmental calibration (protecting the physician's long-term diagnostic calibration). Seven class-specific safeguards were designed, each mapped to the bias pathway it interrupts: a mandatory override checklist and delayed AI display for Class A (Lethal Diagnosis Miss), mechanism verification prompts and conduction analysis disclaimers for Class B (Mechanism Blindness), output suppression thresholds and historical performance transparency for Class C (Signal Corruption), and a quantitative output mode replacing binary alerting for Class D (Self-Undermining AI).

The framework is designed to be net-negative on alert burden. Output suppression of the highest-volume overcalls (old MI at 47.6×, nonspecific ischaemia at 2.7×) removes over 10,000 low-value AI statements from the physician's visual field, while the two safeguards that add alerts (the override checklist and mechanism prompts) together generate fewer than 10 targeted prompts per physician shift. Four of the seven safeguards are deployable through ECG machine software configuration or electronic health record display modification alone, requiring no physician training or workflow change, and collectively address five of eight danger zones. The complete safeguard protocols and implementation framework, including a three-phase rollout strategy and feasibility assessment, are provided in the Supplement.

5.3 Limitations

Seven limitations of this study warrant discussion. First, the analysis is based on the GE 12SL, a legacy rule-based algorithm from the mid-2000s. Contemporary deep neural network systems are likely to show different and potentially attenuated error profiles. However, the analytical framework — the mapping from AI error profiles through cognitive bias pathways to compound clinical risk — is system-agnostic, and the 12SL's extreme error rates serve as a stress test that reveals cognitive risk mechanisms at maximum intensity. The taxonomy's classes and harm mechanisms are predicted to persist even if specific signatures change.

Second, the reference standard comprises expert cardiologist annotations, not emergency department clinician interpretations. This means the study measures AI disagreement with optimal human performance rather than with the typical cognitive environment in which the AI output is actually used. An ED clinician comparator study would be expected to reveal additional disagreement patterns, including anchoring and automation bias effects that the current design cannot capture.

Third, the study measures disagreement, not patient harm. No outcome linkage is available. The Compound Risk Hypothesis is validated through projected performance comparisons using literature-derived parameters, not prospective measurement of actual clinical decisions.

Fourth, the dataset (PTB-XL+) is drawn from a single institution with a predominantly European (German) population. ECG morphology varies with ethnicity, age, and body habitus; error profiles may differ in other populations.

Fifth, the study is retrospective. No physician cognitive responses were directly measured. The bias pathway activations are inferred from the established cognitive bias literature applied to the empirical error profiles, not observed in real-time physician decision-making.

Sixth, the binary per-concept comparison framework may overcount disagreements in borderline cases where both AI and human are uncertain but express uncertainty differently.

Seventh, the safeguard framework is theoretically grounded and evidence-informed but empirically untested. The projected effectiveness of each safeguard is based on cognitive bias theory and analogies to interventions in other clinical decision support domains and requires prospective validation.

5.4 Future Research

Three research directions are prioritised. First, prospective physician testing through randomised simulation in which emergency physicians interpret danger-zone ECGs under three conditions — ECG alone, ECG with unmodified AI output, and ECG with safeguard-modified AI output — would directly validate the Compound Risk Hypothesis and test safeguard effectiveness. Second, modern AI system profiling applying the analytical framework developed here to contemporary deep neural network ECG systems using the same or comparable dual-label datasets would determine whether the identified danger classes persist, attenuate, or transform with current-generation algorithms. Third, outcome-linked analysis connecting AI–human disagreement patterns to patient outcomes (catheterisation laboratory activation, troponin trajectories, disposition decisions, mortality) would move the field from disagreement characterisation to harm quantification.

6. Conclusions

Analysis of 21,799 dual-labelled ECGs across 54 emergency-critical diagnoses revealed systematic AI–human disagreement affecting 81% of ECGs, with a mean of 5.7 concurrent errors per affected tracing. These errors were not random but organised into eight named disagreement signatures, each activating distinct cognitive bias pathways in the receiving physician. The resulting Cognitive Risk Taxonomy classifies these signatures into four danger classes — Lethal Diagnosis Miss, Mechanism Blindness, Signal Corruption, and Self-Undermining AI — with seven classified as CRITICAL and one as HIGH risk.

The Compound Risk Hypothesis was validated across all eight danger zones: in every case, the AI–physician combination was projected to perform worse than the physician alone, because the AI's errors are directionally aligned with the physician's cognitive vulnerabilities. Three harm mechanisms were identified — Direct Suppression, Capability Destruction, and Environmental Contamination — the last of which may persist even after the AI system is removed through bias inheritance.

Two findings have immediate translational implications. First, diagnostic substitution profiles demonstrate that AI misses are not silent omissions but active reframings that provide competing diagnostic narratives, transforming correctable omission errors into resistant commission errors.

Second, the AI's binary confidence architecture, in which 81% of overcalls are delivered at maximum confidence with no hedging, creates the strongest possible anchoring signal for every error, eliminating the confidence gradient that physicians need to calibrate trust.

A seven-component safeguard framework was developed, mapped to the taxonomy's class structure and designed to be net-negative on alert burden. Four of seven safeguards are deployable through software configuration alone.

The taxonomy framework is system-agnostic. While the specific signatures documented here reflect the GE 12SL algorithm, the analytical pipeline — mapping AI error profiles through cognitive bias pathways to compound clinical risk — is designed for reapplication to any AI system with dual-label validation data. As AI-assisted ECG interpretation becomes standard emergency practice, prospective validation of the Compound Risk Hypothesis and safeguard effectiveness through randomised physician simulation studies represents the essential next step.

7. Acknowledgments

This is an Artificial Intelligence-Assisted Research (AIAR); the authors deployed AI tools to increase efficiency and optimize research outcomes. Such usage of AI aligns with the principles of scientific efficiency and effectiveness; the authors have chosen to harness the capacity of contemporary tools. The authors declare no conflict of interest. **Dr. Al-Baraa Dabaliz** (MBBS PGCMedEd, Specialty Doctor Emergency Medicine) and **Dr. Al-Awwab Dabaliz** (MBBS, Cardiac Anesthesiologist and Intensivist) contributed to the medical research and its execution, while **Dr. Mosab Hawarey** (PhD in Geodetic Engineering) contributed to the non-medical research and its execution. **All authors contributed equally to this study.** This study received no external funding.

8. References

1. Bond, Raymond R., Novotny, Tomas, Andrsova, Ivana, Koc, Libor, Sisakova, Maria, Finlay, Dewar, & Guldenring, Daniel. (2018). Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of Electrocardiology*, *51*(1), 6–11. https://doi.org/10.1016/j.jelectrocard.2017.08.007
2. Croskerry, Pat. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, *84*(8), 1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703
3. Croskerry, Pat. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*(8), 775–780. https://doi.org/10.1097/00001888-200308000-00003
4. Gaube, Susanne, Suresh, Harini, Raber, Matthias, Meer, Ankur, Berkowitz, Seth A., Arcas, Blaise Agüera y, Lermer, Eva, Sabelman, Eric E., Ghassemi, Marzyeh, & Colak, Ethna. (2021). Do as AI say: Susceptibility in deployment of clinical decision-making tools. *NPJ Digital Medicine*, *4*(1), 31. https://doi.org/10.1038/s41746-021-00385-9
5. Goddard, Kate, Roudsari, Abdul, & Wyatt, Jeremy C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American*

Medical Informatics Association, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

6. Graber, Mark L. (2013). The incidence of diagnostic error in medicine. *BMJ Quality & Safety*, *22*(Suppl 2), ii21–ii27. https://doi.org/10.1136/bmjqs-2012-001615
7. O'Gara, Patrick T., Kushner, Frederick G., Ascheim, Deborah D., Casey, Donald E., Jr., Chung, Mina K., de Lemos, James A., Ettinger, Steven M., Fang, James C., Fesmire, Francis M., Franklin, Barry A., Granger, Christopher B., Krumholz, Harlan M., Linderbaum, Jane A., Morrow, David A., Newby, L. Kristin, Ornato, Joseph P., Ou, Narith, Radford, Martha J., Tamis-Holland, Jacqueline E., Tommaso, Carl L., Tracy, Cynthia M., Woo, Y. Joseph, & Zhao, David X. (2013). 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, *61*(4), e78–e140. https://doi.org/10.1016/j.jacc.2012.11.019
8. Parasuraman, Raja, & Manzey, Dietrich H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055
9. Rana, Adnan, Grotts, Jason, Engel, Michael, Wang, Mengyu, & Elmore, Joann G. (2025). Anchoring bias of AI-generated ECG interpretation on emergency physicians. *American Journal of Emergency Medicine*, *89*, 18–24. https://doi.org/10.1016/j.ajem.2024.12.005
10. Schläpfer, Jürg, & Wellens, Hein J. (2017). Computer-interpreted electrocardiograms: Benefits and limitations. *Journal of the American College of Cardiology*, *70*(9), 1183–1192. https://doi.org/10.1016/j.jacc.2017.07.723
11. Smith, Stephen W., Rapin, Jérémy, Li, Jian, Thaler, Andrew, Alquier, Pierre, & Elhadad, Noémie. (2016). A deep neural network for real-time detection of acute ST-elevation myocardial infarction. *Journal of Electrocardiology*, *49*(6), 838–843. https://doi.org/10.1016/j.jelectrocard.2016.09.082
12. Strodthoff, Nils, Temme, Patrick, Soni, Aniruddh, Wilhelm, Thorsten, Bischoff, Monja, & Wagner, Patrick. (2023). PTB-XL+, a comprehensive electrocardiographic feature dataset. *Scientific Data*, *10*(1), 279. https://doi.org/10.1038/s41597-023-02153-8
13. Vicente, Lucía, & Matute, Helena. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. https://doi.org/10.1038/s41598-023-42384-8
14. Viskin, Sami, Rosovski, Urszula, Sands, Andrew J., Chen, Elizabeth, Kistler, Peter M., Kalman, Jonathan M., Chavez, Luis R., Torres, Pablo I., Cruz, Fatima E., Centurión, Osmar A., Fujiki, Akira, Maury, Philippe, Chen, Xiaoping, Krahn, Andrew D., Roithinger, Franz, Zhang, Li, Vincent, G. Michael, & Zeltser, Dorit. (2005). Inaccurate electrocardiographic interpretation of long QT: The majority of physicians cannot recognize a long QT when they see one. *Heart Rhythm*, *2*(6), 569–574. https://doi.org/10.1016/j.hrthm.2005.02.011
15. Wagner, Patrick, Strodthoff, Nils, Bousseljot, Ralf-Dieter, Samek, Wojciech, & Schaeffter, Tobias. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, *7*(1), 154. https://doi.org/10.1038/s41597-020-0495-6

Supplementary Materials

Table S1. Disagreement Landscape by ER-Critical Category

Proportion of non-trivial comparisons classified as AI overcall, AI miss, or positive agreement, stratified by the nine ER-critical diagnostic categories.

Category	n Records	AI Overcall (%)	AI Miss (%)	Agreement (%)
Ischemia/ACS Spectrum	32,069	37.3	54.1	8.6
MI Localization	27,448	32.8	56.8	10.4
Conduction Blocks	20,723	34.2	58.3	7.5
Life-Threatening Arrhythmias	13,887	55.4	40.1	4.5
Old MI (Overcall Risk)	7,039	89.6	7.2	3.2
Hypertrophy (STEMI mimic)	3,438	31.2	64.8	4.0
Acute MI	921	43.5	49.9	6.6
QT Prolongation	748	84.4	10.0	5.6
Pre-excitation	128	35.9	64.1	0.0

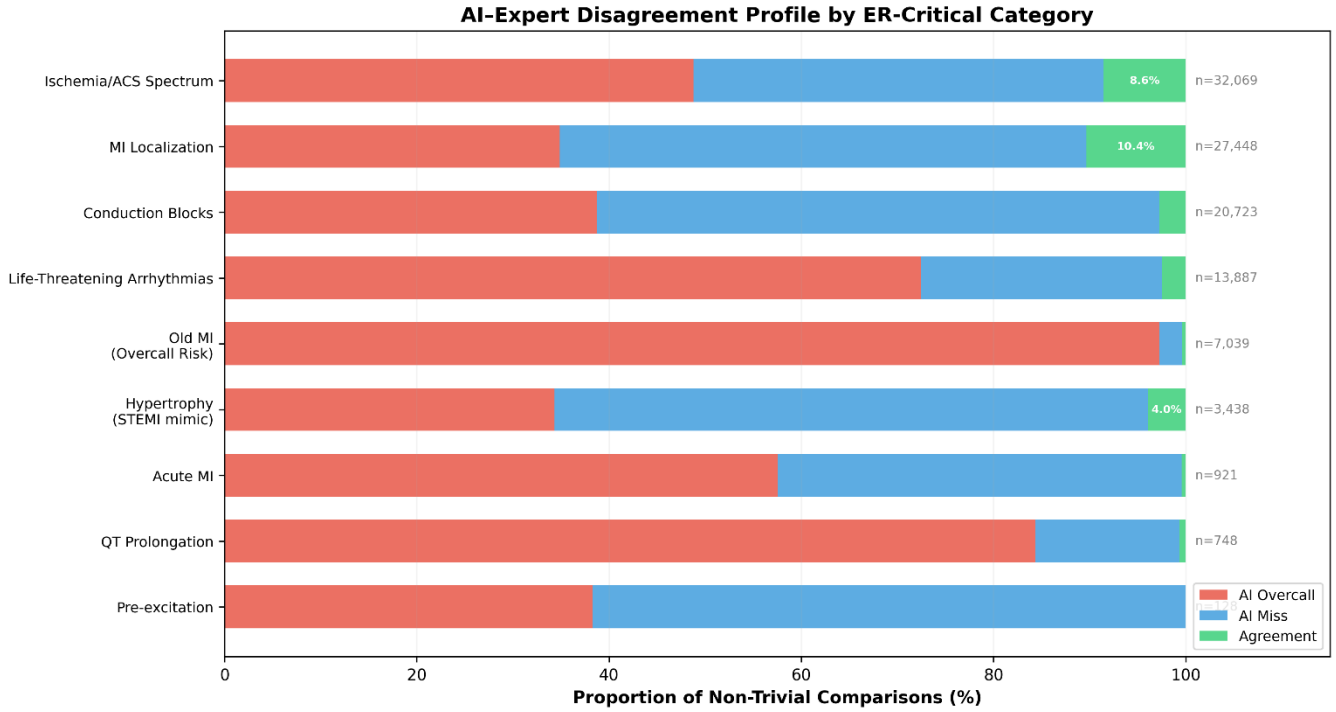

Figure S1. Stacked bar chart of AI–expert disagreement profile by ER-critical category

Table S2. Cohen's κ and McNemar's Test for ER-Critical Concepts
Inter-rater agreement and directional asymmetry for selected emergency-critical SNOMED concepts. Direction: OVERCALL (AI labels more frequently) or MISS (experts label more frequently). Ratio = AI count / Human count.

Concept	κ	McNemar	Direction	Ratio	Signature
Anteroseptal MI on ECG	0.025	<0.001***	MISS	0.2×	SIG-2
Acute inferolateral MI	−0.000	<0.001***	MISS	0.04×	—
ST segment elevation	0.008	<0.001***	OVERCALL	6.6×	—
Old myocardial infarction	0.010	<0.001***	OVERCALL	47.6×	SIG-1
Old inferior MI	0.003	<0.001***	OVERCALL	41.8×	SIG-1
Old anterior MI	0.014	<0.001***	OVERCALL	11.2×	SIG-1
Ventricular tachycardia	−0.001	0.008**	MISS	0.3×	SIG-4
SVT	0.012	<0.001***	OVERCALL	16.2×	SIG-4
Tachyarrhythmia	0.024	<0.001***	OVERCALL	10.6×	SIG-4
Sinus bradycardia	0.261	<0.001***	OVERCALL	5.5×	SIG-3
Bradycardia	0.239	<0.001***	OVERCALL	5.5×	SIG-3
Atrial fibrillation	0.890	<0.001***	MISS	1.1×	— (best κ)
Left bundle branch block	0.283	<0.001***	MISS	0.5×	SIG-6
Right bundle branch block	0.061	<0.001***	MISS	0.6×	—
Complete RBBB	0.361	<0.001***	OVERCALL	1.3×	—
First degree AV block	0.027	<0.001***	OVERCALL	1.6×	—
Left anterior fascicular block	0.164	<0.001***	MISS	0.3×	SIG-7
Left posterior fascicular block	0.057	<0.001***	MISS	0.1×	SIG-7
Lateral ischaemia on ECG	0.010	<0.001***	OVERCALL	9.2×	SIG-5
Nonspecific ST-T changes	0.109	<0.001***	OVERCALL	2.7×	SIG-5
Myocardial ischaemia on ECG	0.217	<0.001***	MISS	0.7×	SIG-5
Ischaemic heart disease	0.321	<0.001***	MISS	0.9×	—
Left ventricular hypertrophy	0.423	<0.001***	MISS	2.9×	—
Prolonged QT interval	0.049	<0.001***	OVERCALL	5.4×	SIG-8

Note: 24 of 54 concepts shown (selected for clinical significance and signature assignment).

Table S3. Complete Error-to-Bias Mapping Matrix
Pathway definitions: A = Overcall → Anchor → Confirm → Close; B = Miss → Omission → Normal anchor; C = Confident-but-wrong → Maximum anchor + System 2 suppression; D = Mass overcall → Prevalence distortion.

Signatures 1–4

Dimension	SIG-1: Old MI Avalanche	SIG-2: Anteroseptal Blind Spot	SIG-3: Bradycardia Inflation	SIG-4: SVT/VT Inversion
Error type	Confident FP	Confident FN	Confident FP (100%)	Simultaneous FP + FN
Magnitude	47.6× OC; ~6,800 FP	0.2× detection; ~2,400 missed	5.5× OC; ~3,400 FP	SVT 16.2× OC + VT 0.3×
Confidence	72% at max	Human 98%; AI silent	100% at max	SVT 100%; Human VT 100%
Pathways	A + C	B + C	A + C	A + B + C (dual)
Primary biases	Anchoring, Automation (commission)	Automation (omission), Anchoring on Normal	Anchoring, Framing	Anchoring, Automation (both), Framing
Secondary biases	Confirmation, Premature closure, Framing	Premature closure, Confirmation, Overconfidence	Premature closure, Base rate neglect	Premature closure, Confirmation
# Biases	5	5	4	5
Novel pattern	Confirmation Cascade	Complacency Loop	Surface-answer trap	Bidirectional Pathway Activation
Compound	768 ECGs (OC + loc miss)	SIG-5 interaction	621 ECGs (OC + block miss)	Tachyarrhythmia amplifier
Score / Tier	8.3 / CRITICAL	9.1 / CRITICAL	9.1 / CRITICAL	8.7 / CRITICAL

Signatures 5–8

Dimension	SIG-5: Ischaemia Noise Floor	SIG-6: LBBB–STEMI Mask	SIG-7: Fascicular Desert	SIG-8: QT Alarm Fatigue
Error type	Mass FP + targeted FN	Confident FN (STEMI-equiv)	Systematic blindness	Confident FP → desensit.
Magnitude	Lat isch 9.2×; NST 2.7×	0.5× detection; ~2,300 miss	LAFB 0.3×, LPFB 0.1×; ~3,500	5.4× OC; 631 FP
Confidence	Mixed	Human 97–98%; AI silent	Human 97–98%; AI silent	100% at max
Pathways	A → future B	B + C	B + C	A → future B
Primary biases	Availability, Algorithm aversion	Automation (omission), Premature closure	Automation (omission), Anchoring on axis	Availability, Algorithm aversion
Secondary biases	Base rate neglect, Desensitisation	Anchoring on Normal, Overconfidence	Overconfidence in AI (axis)	Base rate neglect, Overconfidence own QT
# Biases	4	4	3	4
Novel pattern	Cry-wolf desensitisation	Clinical Pathway Interruption	Delegation Trap	Self-Undermining Loop
Compound	321 ECGs (nonspec OC + miss)	SIG-7 co-occurrence	AV block overcall	Wasted potential
Score / Tier	8.4 / CRITICAL	8.9 / CRITICAL	8.7 / CRITICAL	7.1 / HIGH

Table S4. ER-Critical Concept Selection
54 concepts included from 93 informative SNOMED CT concepts present in both AI and human label sources. Selection based on five criteria: time-critical diagnosis, potential for patient harm if missed, actionable in the emergency department, sufficient sample size (≥10 records), and relevance to established emergency medicine protocols.

Exclusion Summary (39 concepts)

Reason	n	Examples
Non-emergency / benign finding	22	Right atrial enlargement, Premature atrial contraction, Sinus arrhythmia, Nonspecific conduction delay
Not time-critical ER diagnosis	12	Incomplete LBBB, Incomplete RBBB, Disorder of right atrium, Right ventricular abnormality
Insufficient sample size	5	Rare acute MI subtypes with <10 cases

Included Concepts by Category (54 concepts)

Category	Concepts	Signature	Key Concepts			
Acute MI	8	SIG-2	Anteroseptal MI	Inferolateral MI	ST elevation	Posterior MI
Old MI (Overcall Risk)	3	SIG-1	Old MI (47.6×)	Old inferior MI	Old anterior MI	
MI Localization	8	—	Anterior MI on ECG	Lateral MI on ECG	Inferior MI on ECG	+ 5 others
Ischemia/ACS	10	SIG-5	Lateral ischaemia	NST changes	Myocardial ischaemia	+ 7 others
Arrhythmias	10	SIG-3, SIG-4	VT, SVT, Bradycardia	Atrial fibrillation	Tachyarrhythmia	+ 5 others
Conduction Blocks	11	SIG-6, SIG-7	LBBB, RBBB, LAFB	LPFB, 1st deg AVB	Complete AVB	+ 5 others
Hypertrophy	2	—	LVH	RVH		
Pre-excitation	1	—	WPW / Pre-excitation			
QT Prolongation	1	SIG-8	Prolonged QT			

Table S5. Safeguard Protocols and Implementation Framework

S5.1 Safeguard Summary Matrix

ID	Safeguard	Class	Time Cost	Alert Burden	Feasibility	Bias Target	Phase
A1	Mandatory Override Checklist	A	+10–20s	Zero	HIGH	Automation (omission)	III
A2	Delayed AI Display	A	+5–10s	Zero	HIGH	Anchoring, Framing	II
B1	Mechanism Verification Prompt	B	+0s	+1/trigger	MODERATE	Premature closure	II
B2	Conduction Disclaimer	B	+0s	Zero (static)	HIGH	Delegation trap	I
C1	Output Suppression	C	−2s	Negative	VERY HIGH	Availability, Anchoring	I
C2	Performance Badges	C	+0s	Zero (ambient)	HIGH	Overconfidence	I
D1	Quantitative QT Output	D	+0s	Negative	VERY HIGH	Alarm fatigue	I

S5.2 Detailed Safeguard Protocols

Safeguard A1: Mandatory Override Checklist (Class A — Lethal Diagnosis Miss)
Targets: SIG-2, SIG-4, SIG-6. *Layer:* 3 (Omission Detection). *Trigger:* Patient presenting with ACS/arrhythmia-spectrum symptoms.
Before finalising ECG interpretation, physician documents brief assessment of three items: (1) V1–V4 ST-segment assessment (targets SIG-2), (2) Wide-complex tachycardia differentiation (targets SIG-4), (3) Bundle branch block morphology (targets SIG-6). Each requires a documented finding, not a checkbox. This cognitive forcing strategy (Croskerry, 2003) forces System 2 engagement on the three highest-mortality AI blind spots.

Safeguard A2: Delayed AI Display — Read-Then-Reveal (Class A)
Targets: SIG-2, SIG-4, SIG-6. *Layer:* 1 (Output Restructuring).
ECG displayed without AI interpretation. Physician records abbreviated initial impression (rhythm, rate, axis, ST assessment). AI interpretation then revealed. Discrepancies flagged for reconciliation. Eliminates anchoring by ensuring physician's System 1 operates on raw ECG before encountering the AI frame (Bond et al., 2018; Gaube et al., 2021).

Safeguard B1: Mechanism Verification Prompt (Class B — Mechanism Blindness)
Targets: SIG-3, SIG-7. *Layer:* 2 (Forced Analytical Engagement).
When AI flags bradycardia: "Bradycardia detected. ⚠ Verify conduction mechanism: assess PR interval, P:QRS ratio, and AV relationship." When AI reports axis findings: "⚠ Fascicular block assessment: confirm axis evaluation and QRS morphology." Targets premature closure by redirecting attention from surface finding to mechanism.

Safeguard B2: Conduction Analysis Disclaimer (Class B)
Static footer on AI conduction analysis: "Known limitation: This algorithm has reduced sensitivity for fascicular blocks and left bundle branch block. Independent physician assessment of axis deviation and QRS morphology is recommended." Prevents delegation to a system that cannot perform the task. Static to avoid alert fatigue.

Safeguard C1: Output Suppression Threshold (Class C — Signal Corruption)
Suppress from primary display: Old MI (47.6× OC), Old inferior MI (41.8×), Old anterior MI (11.2×) — moved to expandable "AI detail" section. Flag with qualifier: Nonspecific ST-T abnormality (2.7×), Lateral ischaemia (9.2×) displayed with "Low-confidence AI finding — verify against clinical context." Removes highest-volume, lowest-PPV outputs from physician's visual field. Net time saving of ~2 seconds per ECG.

Safeguard C2: Historical Performance Transparency (Class C)
Each AI diagnostic statement annotated with calibration badge: ◍ High agreement (>20% concordance), ◍ Low agreement (<20%), ◍ Very low agreement (<5%). Transforms opaque AI confidence into transparent track record, enabling diagnosis-specific calibrated trust (Parasuraman and Manzey, 2010).

Safeguard D1: Quantitative QT Output Mode (Class D — Self-Undermining AI)
Replace binary "Prolonged QT interval" alert with: "QTc: [measured value] ms (Reference: <450 ms male, <460 ms female). Clinical assessment recommended for values >500 ms." Eliminates 5.4× false alarm rate while retaining AI's genuine computational contribution. Physician applies clinical judgment to determine significance.

S5.3 Phased Implementation Strategy

Phase I — Passive (Immediate): C1, C2, B2, D1. Software configuration only. No training. Addresses 5 of 8 danger zones.
Phase II — Active (3–6 months): A2, B1. Requires orientation sessions. Minor workflow modification.
Phase III — Protocol (6–12 months): A1. Department-level adoption. Pilot in single ED first.

S5.4 Net Alert Burden

The framework is net-negative on alert burden. C1 and D1 remove >10,000 low-value AI outputs. A1 and B1 add <10 targeted prompts per physician shift. Ratio of alerts removed to alerts added: approximately 1000:1.