

Risk-Based Regulatory Boundaries for AI Systems: A Model for Differentiated Governance Across Domains

Song LI

Heilongjiang University, Harbin, China · 747981128@qq.com

Abstract

Risk-based regulation has become the dominant paradigm for governing artificial intelligence, yet how to draw the boundary between “high-risk” and “not-high-risk” systems remains under-theorized. Existing approaches rely on enumerated lists or vague criteria providing insufficient guidance for classification decisions involving novel technologies. This article proposes the **Risk-Based Boundary Model (RBM)**, a three-dimensional framework assessing AI systems along harm potential, decisional autonomy, and contextual vulnerability. Through systematic comparative analysis of the EU AI Act, U.S. sector-specific approaches, and the UK pro-innovation model, we identify both the coherence and systematic gaps of current boundary-setting practices. The RBM provides principled decision rules for classification, domain-specific boundary criteria for healthcare, criminal justice, employment, and education, and audit trigger conditions converting static classification into dynamic regulatory monitoring. Unlike existing frameworks offering risk tiers, sector-specific standards, or governance principles, the RBM functions as a boundary-drawing tool enabling consistent, defensible classification of any AI system, including novel applications not foreseen at the time of legislation. We argue that effective risk-based governance requires not merely risk *tiers* but risk *boundaries*—administrable criteria enabling consistent classification across evolving technological landscapes.

Keywords risk-based regulation, regulatory boundaries, high-risk AI, algorithmic governance, classification framework, EU AI Act, comparative regulation

1 Introduction

1.1 The Boundary Problem in AI Governance

The phrase “high-risk artificial intelligence system” has become perhaps the most consequential regulatory category of the algorithmic age. Systems classified as high-risk face conformity assessments, registration obligations, ongoing monitoring, and enhanced liability exposure. Systems falling outside this classification face minimal or no regulation. The boundary between these categories determines where governance resources concentrate and which algorithmic systems operate in relatively unregulated space.

Yet this boundary—what precisely makes an AI system “high-risk”—has received surprisingly little systematic attention in the governance literature. Despite the proliferation of AI ethics guidelines worldwide (Jobin et al., 2019), the EU AI Act, the most comprehensive algorithmic governance legislation globally, relies primarily on enumerated lists of high-risk application areas (Annex III) supplemented by general criteria that provide limited guidance for classification decisions (European Parliament, 2024). U.S. approaches fragment risk definitions across sector-specific regulatory domains without unified criteria (Engler, 2022). International standards endorse “risk-based approaches” without specifying how risks should be assessed or boundaries drawn (OECD, 2019).

This definitional inadequacy creates three governance problems. First, **classification inconsistency**: functionally similar systems receive different regulatory treatment depending on how applications are framed or which jurisdictional regime applies (Veale & Borgesius, 2021). Second, **regulatory lag**: list-based approaches cannot address novel applications not anticipated when lists were compiled, producing gaps that widen as the technology evolves (Collingridge, 1980). Third, **strategic boundary manipulation**: regulated entities may structure activities to avoid high-risk classification while maintaining essentially similar functionality—a problem well documented in financial regulation (Black, 2008) and now emerging in AI governance (Hacker, 2023).

1.2 The Risk-Based Boundary Model

This article proposes the **Risk-Based Boundary Model (RBM)** as a systematic framework for addressing the boundary problem. The framework comprises three analytically independent dimensions:

1. **Harm Potential**: the severity, reversibility, and scope of possible adverse outcomes
2. **Decisional Autonomy**: the degree to which human judgment is structurally displaced by algorithmic processes
3. **Contextual Vulnerability**: characteristics of affected populations and decision environments that amplify or attenuate risk

The RBM draws on three decades of risk regulation theory (Hood et al., 2001; Baldwin et al., 2011; Black & Baldwin, 2010) and adapts it to the specific challenges of algorithmic governance—opacity, adaptiveness, and scale effects that distinguish AI systems from conventional regulated products. A terminological note: throughout this article, *boundary* refers to an administrable classification condition—a structured set of criteria that triggers a change in regulatory treatment—rather than a simple numerical threshold or cut-off score.

The article makes three distinct contributions. First, it shifts the classificatory logic from enumerated high-risk lists to *dimensional boundary criteria* with explicit decision rules—producing a model that regulators can apply to any AI system, including those not foreseen at the time of legislation. Second, it uses the RBM as a comparative diagnostic tool, systematically revealing how the EU, US, and UK diverge on boundary consistency, adaptation speed, and gaming resistance. Third, it operationalizes the framework through audit trigger conditions and domain-specific boundary rules, connecting abstract classification logic to the concrete regulatory question of *when* re-assessment is warranted and *what events* should compel mandatory review. No prior framework combines analytically independent harm, autonomy, and vulnerability dimensions with explicit, multi-rule classification logic designed as a generalizable boundary-drawing instrument.

1.3 Negative Scoping

Three clarifications delimit the scope of this article. First, the RBM addresses the *classification* of AI systems for regulatory purposes, not the substance of regulatory obligations once classification is made. Questions about what requirements high-risk systems should face—transparency, auditability, human oversight—are addressed in the responsibility allocation literature (Doshi-Velez & Kim, 2017; Selbst, 2021) and are beyond the present scope. Second, the article focuses on *regulatory boundaries* as institutional design choices, not on risk as a social construct or epistemic concept. The sociological question of how risk perceptions are formed (Beck, 1992) is analytically distinct from the institutional question of where regulatory thresholds should be set. Third, the RBM addresses systems that meet the definitional threshold of “AI systems”—those exhibiting autonomous inference from data.

1.4 Article Structure

Section 2 reviews the theoretical foundations of risk-based regulation and their application to AI governance. Section 3 introduces the RBM framework, specifying its three dimensions and decision rules. Section 4 analyzes the policy instruments corresponding to each risk tier. Section 5 provides systematic comparative analysis of the EU, U.S., and UK approaches through the RBM lens. Section 6 discusses institutional implications for enforcement and adaptation. Section 7 concludes.

2 Risk-Based Regulation: Theoretical Foundations

2.1 From Command-and-Control to Risk Proportionality

Risk-based regulation allocates governance resources in proportion to harm magnitude and probability—a principle Black (2005) terms “regulatory proportionality.” Hood, Rothstein, and Baldwin (2001) distinguish three components of any risk regulation regime: *information-gathering*, *standard-setting*, and *behavior modification*. The boundary question—what makes a system “high-risk”—is fundamentally a standard-setting problem: it determines the threshold above which intensive governance applies.

However, risk-based approaches carry their own risks. Classification criteria may ossify; bright-line thresholds invite gaming; regulators may invest disproportionate resources in policing the boundary itself (Rothstein et al., 2006). These meta-regulatory risks are particularly acute in AI governance, where the technology evolves faster than classification systems can adapt.

2.2 Risk Regulation Applied to AI

The transposition of risk-based regulation to AI governance introduces distinctive challenges that the RBM must address. AI systems function as autonomous agents embedded in complex social environments (Rahwan et al., 2019), creating governance problems that differ structurally from traditional regulated products.

Opacity. Unlike traditional regulated products, many AI systems operate as epistemic “black boxes” where the relationship between inputs and outputs is not fully transparent, even to developers (Burrell, 2016), reinforcing arguments for inherently interpretable models (Rudin, 2019). This opacity means that risk assessment cannot rely solely on design specifications; it must also consider operational characteristics and deployment contexts.

Adaptiveness. AI systems, particularly those incorporating continuous learning, may change their behavior after deployment—a property that Collingridge (1980) identified as the fundamental challenge of technology governance. A system classified as low-risk at deployment may drift toward higher-risk behavior through data accumulation, fine-tuning, or adversarial manipulation (Bommasani et al., 2021). Static classification systems cannot capture this dynamism.

Scale and Interconnection. Foundation models and general-purpose AI (GPAI) systems introduce systemic risk—the possibility that failure in a widely deployed model cascades across multiple downstream applications (Novelli et al., 2024). This creates a class of risk that transcends individual application assessment and requires architecture-level evaluation.

Dual-Use and Context-Dependence. The same algorithmic capability may be benign in one context and harmful in another. A facial recognition system used for unlocking a personal device poses different risks than the same technology deployed for mass surveillance (Kaminski, 2019). Risk classification must therefore attend to deployment context, not merely technical capability.

2.3 The Limits of List-Based Approaches

The EU AI Act’s primary mechanism for identifying high-risk systems is enumeration: Annex III lists specific application areas deemed high-risk (European Parliament, 2024). This approach offers legal certainty but suffers from three structural limitations.

First, lists are **backward-looking**: they reflect risks recognized at the time of legislation, not risks that emerge subsequently. Veale and Borgesius (2021) note that the Commission’s power to update Annex III through delegated acts provides some flexibility, but the amendment process introduces significant delay.

Second, lists create **framing incentives**: systems performing essentially identical functions may be classified differently depending on how their “intended purpose” is described. Smuha (2021) observes that the distinction between a “recruitment tool” (high-risk) and a “talent analytics platform” (potentially not listed) may depend more on marketing than on functional characteristics.

Third, lists cannot address **convergent technologies**: systems that combine capabilities from multiple listed and non-listed domains in ways that produce emergent risks not captured by any single category. Renda (2023) argues that this is particularly problematic for GPAI systems, which resist application-specific classification precisely because they are general-purpose.

These limitations motivate the RBM’s principled approach: classification based on dimensional criteria rather than enumerated applications.

3 Theoretical Framework: The Risk-Based Boundary Model

3.1 Framework Overview

The Risk-Based Boundary Model (RBM) provides a systematic approach to defining regulatory boundaries through three dimensions that capture the structural determinants of algorithmic risk. Each dimension is analytically independent—a system may score high on one dimension and low on another—and each is operationalized through observable indicators rather than subjective assessments.

A clarification on dimensional scope is warranted. Other commonly cited risk factors—including scale, opacity, and adaptiveness—are treated within the RBM as *modifiers* of harm potential or decisional autonomy rather than as independent dimensions, to avoid dimensional proliferation that would undermine administrability.

Table 1: Core Definitional Concepts of the Risk-Based Boundary Model

Concept	Definition	Positive Test	Negative Test
Regulatory Boundary	Threshold condition(s) triggering enhanced governance obligations	System satisfies RBM dimensional criteria at specified level	System falls below all dimensional thresholds
High-Risk System	AI system scoring high on ≥ 2 RBM dimensions or critical on any single dimension	Meets multi-dimensional or critical single-dimension threshold	Low scores across all three dimensions
Risk Tier	Category of regulatory intensity determined by aggregate RBM assessment	Assigned through structured dimensional analysis	Cannot be determined without RBM evaluation
Boundary Manipulation	Strategic restructuring of AI deployment to avoid high-risk classification without reducing actual risk	Formal classification changes; substantive risk unchanged	Classification and risk co-vary appropriately

Specifically, scale amplifies harm potential (§3.2, Scope sub-dimension); opacity reduces the effectiveness of human review (§3.3, structural conditions); adaptiveness triggers re-classification through audit mechanisms (§6.2).

3.2 Dimension One: Harm Potential

The first dimension assesses the potential for adverse outcomes from algorithmic operations. Following Sunstein’s (2002) distinction between quantifiable risks and Knightian uncertainties, harm potential comprises three sub-dimensions assessed through observable indicators.

Severity measures the magnitude of harm if adverse outcomes materialize. Drawing on fundamental rights frameworks (FRA, 2020), severity ranges from minor inconvenience through economic loss, discrimination, and rights violations to threats to life and physical integrity.

Reversibility captures whether harms, once realized, can be undone. This dimension is critical because irreversible harms justify precautionary regulatory intervention even at lower probability levels (Sunstein, 2005). Criminal records generated by biased risk assessment tools, for instance, create path dependencies that persist long after the algorithmic error is corrected.

Scope measures the number and concentration of affected individuals. Mass-scale deployment amplifies aggregate harm even when per-individual harm is moderate—a consideration that distinguishes AI regulation from traditional product safety, where harm typically affects individual users (Novelli et al., 2024).

Table 2: Dimension 1—Harm Potential Assessment Matrix

Level	Severity	Reversibility	Scope	Illustrative Applications
Critical	Life, physical integrity, liberty	Irreversible or effectively permanent	Any	Autonomous weapons targeting; lethal medical device decisions
High	Fundamental rights; significant economic impact	Difficult to reverse; substantial remediation costs	Group-scale or broader	Employment screening; credit scoring; criminal risk assessment
Medium	Moderate economic or informational impact	Reversible with reasonable effort	Variable	Content recommendation; dynamic pricing; targeted advertising
Low	Minor inconvenience or preference frustration	Easily reversible; trivial remediation	Individual	Spam filtering; personal media recommendation

3.3 Dimension Two: Decisional Autonomy

The second dimension assesses the degree to which human judgment is structurally displaced by algorithmic processes. This dimension captures a concern distinct from harm potential: even when potential harms are moderate, complete displacement of human decision-making erodes accountability and reduces opportunities for error correction (Yeung, 2018).

The critical insight is that formal “human-in-the-loop” arrangements may provide false assurance where structural conditions undermine meaningful human oversight. Green and Chen (2019) demonstrate experimentally that human decision-makers presented with algorithmic recommendations exhibit “automation bias”—systematically deferring to machine outputs even when those outputs are incorrect.

Table 3: Dimension 2—Decisional Autonomy Levels

Level	System Role	Human Capacity	Structural Conditions	Examples
Full Autonomy	System decides and executes without human involvement	None or purely nominal	No review mechanism; no override capability	Automated trading execution; autonomous vehicle operation
High Autonomy	System decides; human implements without substantive review	Implementation only; no realistic capacity to override	High volume; time pressure; asymmetric expertise	Automated benefit denial with appeal as sole recourse
Constrained Review	System recommends; human formally decides under structural constraints	Formally present but practically constrained	High caseload; algorithmic authority; information asymmetry	Judicial risk scores with heavy dockets; loan decisions with time quotas
Genuine Review	System provides information; human independently evaluates	Full decisional authority; adequate time and information	Low volume; transparent methodology; domain expertise	Decision support tools with explainable outputs

3.4 Dimension Three: Contextual Vulnerability

The third dimension assesses characteristics of the populations affected by algorithmic decisions and the environments in which those decisions occur. Vulnerability is not an intrinsic property of individuals but a structural feature of decision contexts—it arises from power asymmetries, information deficits, and dependency relationships that prevent affected parties from protecting their own interests (Fineman, 2008).

Power Asymmetry captures relationships where the party subject to algorithmic decision-making cannot effectively negotiate, exit, or resist. **Information Asymmetry** captures situations where affected parties lack knowledge of how algorithmic systems operate (Burrell, 2016). **Dependency** captures contexts where affected parties cannot avoid algorithmic decision-making. **Protected Characteristics** capture contexts where decisions disproportionately affect groups subject to historical discrimination (Barocas & Selbst, 2016).

Table 4: Dimension 3—Contextual Vulnerability Levels

Level	Power Asymmetry	Information Deficit	Dependency	Protected Groups
High	Substantial; no meaningful exit option	Near-total opacity; no explanation mechanisms	Essential service; no alternative	Decision directly affects protected characteristics
Medium	Moderate; limited alternatives	Partial transparency; some explanation available	Important but substitutable	Indirect or incidental effects on protected groups
Low	Relative parity; multiple options	Substantial transparency; informed consent viable	Discretionary; alternatives readily available	No differential impact on protected groups

3.5 Classification Decision Rules

The RBM specifies four decision rules for translating dimensional assessments into regulatory classifications:

Rule 1—Critical Threshold: A system scoring *critical* on harm potential warrants the highest regulatory tier regardless of other dimensional scores. This reflects the precautionary principle that irreversible catastrophic harms justify intensive regulation even under uncertainty (Sunstein, 2005).

Rule 2—Multi-Dimensional Threshold: A system scoring *high* on at least two of three RBM dimensions qualifies as high-risk. This threshold ensures that systems must present *structural* risk—reflected across multiple independent dimensions.

Rule 3—Aggregate Assessment with Justification: Where dimensional scores produce mixed results, classification requires documented justification addressing each dimension.

Rule 4—Precautionary Default: Uncertainty about dimensional scores resolves toward higher classification. This asymmetric error preference reflects the judgment that under-regulation of genuinely risky systems is more costly than temporary over-regulation of systems later reclassified downward (Sunstein, 2005).

Two walkthrough examples illustrate how the decision rules operate in practice:

Example A—Medical Diagnostic Chatbot. A healthcare chatbot that triages symptoms and recommends treatment pathways in a resource-constrained public health setting. Harm potential: **High** (diagnostic error may delay treatment; partially irreversible). Decisional autonomy: **High** (patients interact without clinician mediation). Contextual vulnerability: **High** (significant patient-platform power asymmetry; near-total opacity; high dependency). Under Rule 2, the system classifies as **high-risk**.

Example B—Enterprise Inventory Optimization. A fully automated system managing warehouse stock levels for a manufacturing firm. Harm potential: **Low** (cost inefficiency and temporary stock-outs; readily reversible). Decisional autonomy: **Full**, but this alone does not trigger intensive regulation. Contextual vulnerability: **Low** (B2B context; negotiating parity; multiple alternatives). Under Rule 2, only one dimension scores high; the system classifies as **limited** or **minimal** risk.

Table 5: RBM Risk Classification Matrix

Risk Category	RBM Dimensional Profile	Regulatory Approach	Policy Instruments
Unacceptable	Critical harm + any autonomy + any vulnerability; OR practices fundamentally incompatible with democratic values	Absolute prohibition with narrow exceptions	Ban; criminal sanctions; market withdrawal
High	High on ≥ 2 dimensions; OR critical harm with genuine review	Ex-ante conformity assessment; continuous compliance monitoring	Quality management systems; technical documentation; registration; human oversight requirements
Limited	Medium harm + high autonomy; OR high on single dimension only	Transparency and disclosure obligations	Interaction awareness; content labeling; user notification
Minimal	Low across all dimensions	Voluntary codes of conduct; general consumer protection law	Industry self-regulation; best practice guidelines

3.6 Operational Note

The RBM is not intended as a mechanical scoring algorithm that produces determinate numerical outputs. It is a *structured decision protocol* requiring regulators to explicitly document dimensional assessments and justificatory reasoning for each classification decision. This design reflects a deliberate trade-off: the RBM sacrifices the false precision of fully quantified risk scores in favor of transparency and contestability in classification reasoning.

3.7 Anticipating Objections: Why Boundaries Resist Full Quantification

Three foreseeable objections to the RBM's design warrant anticipatory response.

First, one might argue that a **fully quantified scoring model** would produce more consistent outcomes. The difficulty is that quantification in this domain manufactures precision that the underlying phenomena do not support. Any weighting scheme embeds normative choices that are themselves contestable, and the resulting scores create an illusion of objectivity that obscures the value judgments at stake (Jasanoff, 2016).

Second, one might suggest that **expanding existing lists** would achieve the same effect with less institutional disruption. List expansion addresses the coverage problem but not the structural one: every addition requires the same legislative process, and every new entry invites the same framing and derogation strategies. More fundamentally, lists cannot provide classification criteria for systems that *do not yet exist*.

Third, one might question whether a **procedural standard** provides weaker protection than a **metric standard**. AI governance involves risks that are context-dependent, multi-dimensional, and evolving—characteristics that favor procedural standards requiring structured justification over metric standards that pre-commit to specific quantitative thresholds.

4 Policy Instrument Analysis

4.1 Instruments for Unacceptable Risk: Prohibition as Bright-Line Boundary

The RBM's highest tier employs prohibition—the most restrictive policy instrument available. Following Baldwin et al.'s (2011) taxonomy of regulatory instruments, prohibition is appropriate only where the harm potential is so severe that no degree of risk mitigation renders deployment acceptable.

The EU AI Act identifies four prohibited practices: manipulative subliminal techniques, exploitation of vulnerable individuals, social scoring by public authorities, and untargeted facial recognition in public spaces (European Parliament, 2024, Art. 5). Through the RBM lens, each of these practices scores *critical* on harm potential, *high* on contextual vulnerability, and *high* on decisional autonomy. The three-dimensional

analysis reveals that prohibition is triggered when all three dimensions simultaneously read at maximum intensity—a convergence that signals risks fundamentally incompatible with democratic governance.

4.2 Instruments for High Risk: The Ex-Ante Compliance Architecture

For systems classified as high-risk under the RBM, the appropriate policy instrument regime combines ex-ante conformity assessment with continuous post-market monitoring. This dual-phase architecture reflects the recognition that static pre-deployment assessment is insufficient for adaptive systems (Mökander et al., 2021).

Pre-market obligations include risk management systems requiring providers to implement systematic processes for identifying, analyzing, and mitigating risks throughout the AI lifecycle (Buiten, 2019); data governance ensuring training datasets meet quality criteria addressing representativeness and freedom from systematic bias; technical documentation describing system architecture and performance metrics; and fundamental rights impact assessments (FRIAs) that operationalize the vulnerability dimension of the RBM.

Post-market obligations include automated logging sufficient to enable retrospective analysis of contested decisions, incident reporting within specified timeframes, and periodic re-assessment of risk classification (minimum every three years) incorporating deployment experience and technological evolution.

4.3 Instruments for Limited Risk: Transparency as Minimum Standard

Systems scoring medium on a single dimension face transparency obligations rather than conformity assessment. Users must be informed when interacting with AI systems, and AI-generated content must be identifiable as such. Transparency obligations reduce information asymmetry—a component of contextual vulnerability—and create preconditions for informed consent.

4.4 Instruments for Minimal Risk: Voluntary Self-Regulation

Systems scoring low across all RBM dimensions need not face AI-specific regulation. General consumer protection law, competition law, and data protection requirements provide baseline governance without imposing sector-specific compliance burdens.

5 Comparative Analysis: EU, US, and UK Through the RBM Lens

5.1 Analytical Framework

This section applies the RBM's three dimensions systematically to evaluate the boundary-setting practices of three major jurisdictions.

5.2 European Union: Comprehensive Enumeration, Definitional Gaps

The EU AI Act represents the most developed operationalization of risk-based AI governance (European Parliament, 2024). Applied through the RBM lens, the Act's high-risk categories generally align with systems scoring high on at least two RBM dimensions. However, the RBM reveals three systematic gaps.

Gap 1: Algorithmic content governance. Large-scale content moderation systems score high on autonomy, medium-to-high on harm potential (information ecosystem effects), and high on vulnerability (massive power asymmetry between platform and user). Yet they are not classified as high-risk under Annex III.

Gap 2: General-purpose AI at the application layer. The Act addresses GPAI models at the provider level but provides limited guidance for classifying downstream applications. The RBM's dimensional approach assesses GPAI-derived applications at the deployment layer through the same three-dimensional

Table 6: Cross-Jurisdictional Comparison of Risk Boundary Architecture

Dimension	EU AI Act	US Sector-Specific	UK Pro-Innovation
Risk Tiers	4 tiers (prohibited, high, limited, minimal)	No unified tiers; sector-specific thresholds	5 principles; no formal tiers
Classification Method	Enumerated lists (Annex III) + safety component integration	Existing regulatory authority applied to AI applications	Context-specific regulator assessment
Harm Potential Coverage	Comprehensive via Annex III categories	Strong in healthcare (FDA), finance; gaps in cross-cutting applications	Distributed across sectoral regulators; uneven coverage
Autonomy Assessment	Implicit; Art. 14 requires human oversight without autonomy classification	Variable by sector	Addressed through “accountability” principle
Vulnerability Assessment	Implicit; specific mentions of children, employees, criminal defendants	Strong in employment (EEOC); weaker in other domains	Addressed through “fairness” principle
Adaptation Mechanism	Delegated acts to update Annex III; regulatory sandboxes	Agency rulemaking within existing statutory authority	Regulatory flexibility by design
Enforcement	Fines up to €35M or 7% global turnover	Variable: sector-specific penalties; FTC consent decrees	Under development
Brussels Effect	High: extraterritorial reach by design (Bradford, 2020)	Low: primarily domestic scope	Medium: influence through GPAI safety initiatives

analysis applied to any other AI system. Compute-based thresholds such as the 10^{25} FLOPs benchmark remain useful as *screening indicators* but are insufficient as *boundary criteria* because identical compute levels may produce vastly different risk profiles depending on downstream application.

Gap 3: Derogation gaming. Article 6(3) creates derogations for systems performing “narrow procedural tasks,” even within Annex III domains. The RBM reveals that these derogations assess only one dimension while ignoring autonomy and vulnerability.

5.3 United States: Sectoral Depth, Architectural Gaps

The United States lacks comprehensive federal AI legislation, relying instead on existing regulatory authorities (Engler, 2022). The FDA’s medical device classification system exemplifies strong RBM alignment: Class III devices—those posing the greatest risk to patients—receive the most intensive pre-market scrutiny, including clinical trials and pre-market approval rather than mere notification. Empirical evidence of racial bias in algorithmic healthcare systems (Obermeyer et al., 2019) underscores the critical importance of dimensional assessment in medical AI classification.

In employment, the Equal Employment Opportunity Commission (EEOC) has applied Title VII’s disparate impact framework to algorithmic hiring tools, treating AI-driven candidate screening as a “selection procedure” subject to the Uniform Guidelines on Employee Selection Procedures. This operationalizes the RBM’s vulnerability dimension within a specific sectoral context. Similarly, the Federal Trade Commission (FTC) has pursued enforcement actions against deceptive or unfair algorithmic practices under Section 5 of the FTC Act, effectively policing the harm potential dimension through ex-post enforcement rather than ex-ante classification.

The architectural limitation is that no mechanism integrates these sectoral assessments into a unified classification framework. The FDA, EEOC, and FTC each assess different RBM dimensions within their respective mandates, but cross-cutting AI systems that span multiple sectors—such as a wellness platform combining health monitoring, employment recommendations, and financial scoring—fall into regulatory gaps that no single agency is positioned to address.

5.4 United Kingdom: Principled Flexibility, Boundary Uncertainty

The UK’s pro-innovation approach tasks existing regulators with applying five cross-cutting principles within their respective domains (UK Government, 2023). Through the RBM lens, this approach offers maximal flexibility but minimal boundary specificity. The five principles correspond roughly to RBM dimensions: safety maps onto harm potential; fairness onto vulnerability; accountability onto the autonomy–oversight relationship. But *principles without thresholds do not produce administrable boundaries*. The UK case illustrates a general lesson: principles identify relevant considerations; boundaries specify when those considerations trigger regulatory consequences.

5.5 Comparative Synthesis

The comparative analysis reveals a convergence–divergence pattern:

Table 7: Cross-Jurisdictional Assessment Quality

Assessment Quality	EU	US	UK
Explicit dimensional criteria	Partial (enumeration + vague criteria)	Variable by sector	Minimal (principles only)
Boundary consistency	High within Annex III; gap outside	High within sectors; gaps between them	Low (regulator-dependent)
Adaptation speed	Moderate (delegated acts)	Variable (agency rulemaking)	High (regulatory discretion)
Gaming resistance	Medium (derogation loopholes)	Medium (framing incentives)	High (no bright lines to game)

The RBM provides the principled criteria that each approach partially implements and partially lacks. Crucially, the RBM does not *replace* these existing approaches but *subsumes* them within a unified analytical structure. The EU’s Annex III list can be understood as predetermined boundary points within the RBM’s dimensional space. The US sectoral definitions represent domain-specific operationalizations of one or two RBM dimensions. The UK’s principles identify the relevant considerations that the RBM’s dimensions formalize.

6 Institutional Implications

6.1 Domain-Specific Boundary Rules

The RBM’s three-dimensional framework must be operationalized through domain-specific boundary criteria that translate abstract dimensional assessments into administrable classification decisions.

6.2 Audit Trigger Conditions

Beyond initial classification, the RBM specifies conditions that trigger re-assessment of an AI system’s risk classification during its operational lifecycle.

6.3 The Expertise Gap and Institutional Capacity

The RBM’s effectiveness depends on regulatory institutions possessing the technical capacity to conduct dimensional assessments. Coglianese (2017) observes that performance-based regulation demands regulators capable of measuring outcomes rather than merely prescribing processes. Current regulatory bodies face severe capacity constraints relative to the technical complexity of frontier AI systems (Renda, 2023).

Table 8: Domain-Specific RBM Application

Domain	Primary Risk Factors	Boundary Criteria (High-Risk Threshold)	RBM Dimensional Profile
Healthcare	Diagnostic error; treatment allocation bias; device safety failure	AI used for triage, diagnosis, treatment recommendation, or as safety component in medical devices	Harm: Critical (life/health); Autonomy: High; Vulnerability: High (patient dependency)
Criminal Justice	Sentencing disparity; pretrial detention bias; surveillance overreach	AI used for risk assessment, sentencing input, evidence evaluation, or predictive policing	Harm: Critical (liberty); Autonomy: Constrained review; Vulnerability: High (state power asymmetry)
Employment	Systematic discrimination in hiring; performance evaluation bias	AI used for candidate screening, interview evaluation, promotion/termination decisions	Harm: High (economic opportunity); Autonomy: High; Vulnerability: High (employer–employee asymmetry)
Education	Deterministic tracking; assessment bias; surveillance-driven proctoring	AI used for admissions, grading, performance assessment, or adaptive learning pathway assignment	Harm: High (developmental impact); Autonomy: Medium–High; Vulnerability: High (children; institutional asymmetry)

Table 9: RBM Audit Trigger Framework

Trigger Type	Threshold	Verification Method	Primary Responsibility	Regulatory Response
Impact Anomaly	Outcome disparities across protected groups exceeding baseline by ≥ 2 SD	Post-market monitoring; outcome audit; FRIA	Deployer; Regulator	Mandatory audit; potential suspension
Automation Degree Shift	Transition from human-in-the-loop to autonomous operation	Control interface audit; workflow analysis	Provider/Deployer; Regulator	Re-classification required
Vulnerability Expansion	Deployment to new populations or non-voluntary contexts	Deployment scope monitoring; demographic analysis	Deployer; Regulator	Re-assessment of vulnerability dimension
Performance Degradation	Accuracy/fairness metrics below declared thresholds for ≥ 30 days	Continuous performance monitoring	Provider; Deployer	Enhanced monitoring; corrective action
Incident Accumulation	≥ 3 serious incidents within 12 months, or any single death/irreversible harm	Incident reporting system + root cause assessment	Provider/Deployer; Regulator	Immediate enhanced scrutiny; potential suspension

The RBM addresses this through two institutional design recommendations. First, **centralized technical expertise**: specialized AI audit agencies would benefit from developing and maintaining the technical capacity for dimensional assessment. Second, **co-regulatory standards with public oversight**: while technical standardization by industry bodies can operationalize the RBM’s dimensional criteria, the standard-setting process itself is likely to require public scrutiny to mitigate the risk of regulatory capture (Floridi, 2021).

The capacity gap is particularly acute in the Global South, where AI systems designed in high-income countries are increasingly deployed in high-vulnerability contexts. The RBM’s vulnerability dimension provides a structural safeguard: where power asymmetry, information deficit, and dependency simultaneously read at maximum levels, the framework assigns high-risk classification regardless of the deploying jurisdiction’s domestic regulatory maturity. The RBM thus offers a principled response to concerns about “algorithmic colonialism” (Mohamed et al., 2020)—the deployment of undertested systems in populations excluded from design, training, and governance processes.

6.4 Regulatory Sandboxes and Adaptive Governance

The RBM incorporates regulatory sandboxes as an institutional mechanism for addressing classification uncertainty. Where a system’s dimensional profile does not clearly resolve to a specific risk tier, sandbox environments allow supervised deployment with enhanced monitoring. In the RBM’s architecture, sandboxes function as the empirical complement to Rule 3 (Aggregate Assessment with Justification): where dimensional scores produce mixed or indeterminate results, sandbox evidence provides the observational basis for documented classification decisions.

7 Conclusion

This article has proposed the **Risk-Based Boundary Model (RBM)** as a principled approach to the central definitional challenge in AI governance: where to draw the boundary between systems that warrant intensive oversight and those that do not. AI governance frameworks must account for both risks and opportunities across societal domains (Vinuesa et al., 2020), and the RBM provides the analytical architecture for doing so. Three conclusions follow.

First, **dimensional criteria are likely to prove more robust than enumerated lists**. The RBM subsumes partial approaches within a unified three-dimensional structure—harm potential, decisional autonomy, contextual vulnerability—and adds explicit decision rules that link assessment to classification. Because the RBM assesses structural properties rather than application labels, it applies to novel systems not anticipated at the time of legislation.

Second, **boundaries must be dynamic**. The audit trigger framework (§6.2) addresses what static classification cannot: AI systems that change their risk profile after deployment through automation degree shifts, vulnerability expansion, or performance degradation.

Third, **institutional capacity is likely the binding constraint**. The expertise gap between regulatory ambition and enforcement capacity would benefit from sustained investment in technical audit infrastructure, international cooperation on assessment methodologies, and public oversight of industry-led standardization.

Regulatory bodies may adopt the Risk-Based Boundary Model as a definitional standard for determining which AI systems warrant intensive oversight. The RBM does not resolve all contested questions about appropriate risk thresholds—those remain matters for democratic deliberation. But the Model provides the analytical architecture within which such deliberation can proceed with conceptual clarity rather than ad hoc enumeration.

References

- Baldwin, R., Cave, M., & Lodge, M. (2011). *Understanding regulation: Theory, strategy, and practice* (2nd ed.). Oxford University Press.
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. <https://doi.org/10.15779/Z38BG31>
- Beck, U. (1992). *Risk society: Towards a new modernity*. Sage Publications.
- Black, J. (2005). The emergence of risk-based regulation and the new public risk management in the United Kingdom. *Public Law*, 2005(Autumn), 512–548.
- Black, J. (2008). Forms and paradoxes of principles-based regulation. *Capital Markets Law Journal*, 3(4), 425–457. <https://doi.org/10.1093/cmlj/kmn026>
- Black, J., & Baldwin, R. (2010). Really responsive risk-based regulation. *Law & Policy*, 32(2), 181–213. <https://doi.org/10.1111/j.1467-9930.2010.00318.x>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arber, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.
- Buiten, M. C. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, 10(1), 41–59. <https://doi.org/10.1017/err.2019.8>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Coglianesi, C. (2017). The limits of performance-based regulation. *University of Michigan Journal of Law Reform*, 50(3), 525–563.
- Collingridge, D. (1980). *The social control of technology*. Frances Pinter.

- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Engler, A. (2022). The EU AI Act will have global impact, but a limited Brussels Effect. *Brookings Institution*.
- European Parliament. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.
- Fineman, M. A. (2008). The vulnerable subject: Anchoring equality in the human condition. *Yale Journal of Law & Feminism*, 20(1), 1–23.
- Floridi, L. (2021). The European Legislation on AI: A brief analysis of its philosophical approach. *Philosophy & Technology*, 34(2), 215–222. <https://doi.org/10.1007/s13347-021-00460-9>
- FRA—European Union Agency for Fundamental Rights. (2020). *Getting the future right: Artificial intelligence and fundamental rights*. Publications Office of the European Union.
- Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–24. <https://doi.org/10.1145/3359152>
- Hacker, P. (2023). The European AI liability directives—Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51, 105871. <https://doi.org/10.1016/j.clsr.2023.105871>
- Hood, C., Rothstein, H., & Baldwin, R. (2001). *The government of risk: Understanding risk regulation regimes*. Oxford University Press.
- Jasanoff, S. (2016). *The ethics of invention: Technology and the human future*. W. W. Norton & Company.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kaminski, M. E. (2019). Binary governance: Lessons from the GDPR’s approach to algorithmic accountability. *Southern California Law Review*, 92(6), 1529–1616.
- Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684. <https://doi.org/10.1007/s13347-020-00405-8>
- Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27(4), 44. <https://doi.org/10.1007/s11948-021-00319-4>
- Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). Taking AI risks seriously: A new assessment model for the AI Act. *AI & Society*, 39(5), 2493–2497. <https://doi.org/10.1007/s00146-023-01723-z>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449.
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- Renda, A. (2023). Making the EU AI Act work: The implementation challenge ahead. *CEPS Policy Insights*, 2023–14.
- Rothstein, H., Irving, P., Walden, T., & Yearsley, R. (2006). The risks of risk-based regulation: Insights from the environmental policy domain. *Environment International*, 32(8), 1056–1065. <https://doi.org/10.1016/j.envint.2006.06.008>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

- Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, 35(1), 117–191.
- Smuha, N. A. (2021). From a ‘race to AI’ to a ‘race to AI regulation’: Regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1), 57–84. <https://doi.org/10.1080/17579961.2021.1898300>
- Sunstein, C. R. (2002). *Risk and reason: Safety, law, and the environment*. Cambridge University Press.
- Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle*. Cambridge University Press.
- UK Government. (2023). *A pro-innovation approach to AI regulation*. Department for Science, Innovation and Technology. CP 815.
- Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4), 97–112. <https://doi.org/10.9785/cri-2021-220402>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11, 233. <https://doi.org/10.1038/s41467-019-14108-y>
- Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523. <https://doi.org/10.1111/rego.12158>

Academic Editor *Xin Huang*, cpcl@cpcl.hk