Governing Intelligent Systems: Global Frameworks,Implementation Gaps, and Pathways to Protect Humanity

Munongedzi Mabhoko, mabhokm@clarkson.edu, Clarkson University

Abstract

Over the last decade, governments, standards bodies, and firms have rapidly converged on a shared vocabulary of "trustworthy," "responsible," and "human-centric" AI. Major frameworks now exist at global (UNESCO, OECD, UN, G7), regional (EU AI Act, African Union Continental AI Strategy, China's algorithmic measures), national (US executive actions, sectoral rules), and organizational levels (NIST AI Risk Management Framework, ISO/IEC 42001). Yet empirical evidence from 2023–2025 shows a persistent implementation gap, AI and other intelligent systems diffuse faster than governance structures, and real-world practices frequently fall outside or beneath these frameworks. Surveys report that AI use is now mainstream (≈78% of organizations) while fewer than half of firms have any formal AI governance policy, and only a small minority have deeply implemented controls. Public trust is fragile, global surveys find rising concern about fairness, privacy, and misinformation, especially in advanced economies. This thesis synthesizes recent research on what has been built, how it has actually been implemented across regions (Europe, North America, China, Africa, emerging economies), why gaps persist, how "rogue elements" emerge (shadow AI, disinformation, unregulated uses), and what design changes are needed if governance is to genuinely protect human dignity while preserving the benefits of intelligent systems.

Chapter 1: Introduction

1.1 Background

Intelligent systems including contemporary machine learning models, decision-support algorithms, recommender systems, and large-scale generative models now mediate access to work, credit, healthcare, education, media, and public services. Their rapid diffusion has

outpaced traditional regulatory cycles, prompting a wave of new AI governance instruments at global, regional, and national levels. These include the OECD AI Principles and their implementation reviews , UNESCO's Recommendation on the Ethics of Artificial Intelligence , the European Union's Artificial Intelligence Act (EU AI Act) , the NIST AI Risk Management Framework (AI RMF) and its Generative AI profile, China's layered algorithm and generative AI regulations , and the African Union's Continental AI Strategy .

At the same time, empirical evidence shows that AI adoption has gone mainstream while governance remains partial and uneven. The 2024 Stanford AI Index reports that AI use among large firms has grown dramatically, but most organizations have operationalized only a subset of the risk mitigations they themselves consider relevant . Surveys by Pew Research Center and others show that publics in many countries are more concerned than excited about AI, with fears clustered around privacy, job displacement, manipulation, and loss of human agency. A large fraction of workers now use "shadow AI" unapproved AI tools used outside formal governance structures often with sensitive data [16]. This juxtaposition of formal governance frameworks and persistent distrust suggests a core problem, the presence of frameworks does not guarantee effective, trusted governance.

1.2 Research problem and questions

This thesis addresses the following merged research question:

> **How effective are current AI governance frameworks (such as the EU AI Act, NIST AI RMF, OECD Principles, UNESCO Recommendations, and China's AI regulations) at addressing real-world risks posed by intelligent systems, and to what extent do these frameworks align with public perceptions, fears, and expectations about AI? Additionally, how does this alignment or misalignment influence the overall effectiveness and public trust in AI governance?**

This question has three dimensions:

1. **Effectiveness:** Do existing frameworks materially reduce the real-world risks associated with intelligent systems (e.g. discrimination, opacity, security failures, disinformation, harmful automation)?
2. **Alignment:** Do the priorities, concepts, and remedies embedded in these frameworks correspond to what publics actually fear, expect, and experience?

3. **Interaction:** How does alignment or misalignment between governance and public sentiment shape compliance, shadow AI, and trust?

1.3 Hypothesis

The working hypothesis is:

> Current AI governance frameworks are structurally insufficient because they rely heavily on broad principles, voluntary compliance, and reactive policymaking. While they aim to mitigate risks like bias, opacity, disinformation, and algorithmic harm, they often fail to address the specific concerns that the public actually experiences or fears. This misalignment between governance priorities and public sentiment weakens public trust and reduces compliance, ultimately diminishing the real-world effectiveness of these frameworks. A more effective governance model would integrate enforceable regulatory mechanisms with participatory, human-centered design, grounded in empirical evidence and community input.

1.4 Significance and contributions

1. **Synthesis of current frameworks:** It offers an integrated, cross-regional analysis of major AI governance instruments (EU, US/NIST, OECD, UNESCO, China, African Union), including their scope, mechanisms, and implementation status, drawing on recent policy and academic work .
2. **Effectiveness assessment:** It evaluates the effectiveness of these frameworks using secondary evidence from surveys, sectoral case studies (notably healthcare), and comparative governance analyses.
3. **Alignment analysis:** It develops a conceptual lens for **governance public alignment**, using public opinion data, and shows how misalignment contributes to shadow AI, partial compliance, and contested legitimacy.

1.5 Thesis structure

The remainder of the thesis is organized as follows:

- **Chapter 2** reviews the literature on AI governance frameworks, implementation, and public perceptions.
- **Chapter 3** outlines the methodology: a comparative policy analysis using secondary empirical data and an alignment framework.
- **Chapter 4** presents the findings on framework effectiveness and governance–public alignment across regions.
- **Chapter 5** discusses implications, revisits the hypothesis, and proposes design principles for more effective, human-centered, and participatory intelligent-systems governance.
- A brief **Conclusion** summarizes contributions and future research directions.

Chapter 2 – Literature Review

2.1 Global and regional AI governance frameworks

2.1.1 OECD AI Principles

In 2019 OECD member states adopted the OECD AI Principles, which articulate values-based norms for "trustworthy AI" and policy recommendations for governments . An implementation review in 2023 reports over 1,000 policy initiatives across 70 jurisdictions citing or aligned with these principles, including national strategies, regulatory proposals, and sectoral guidelines [4][6].

The principles emphasize inclusive growth, human-centered values, transparency, robustness, security, and accountability, but they are **non-binding**. Their effectiveness depends on how states translate them into law, institutions, and practice.

2.1.2 UNESCO Recommendation on the Ethics of AI

UNESCO's 2021 Recommendation is the first global standard on AI ethics adopted by all UNESCO member states. It anchors AI governance in human rights, human dignity, and environmental sustainability, and calls for human oversight, fairness, and transparency. Recent updates and implementation discussions stress that states must adopt national policies, impact assessment mechanisms, and capacity-building programs to actualize the Recommendation [5][25].

While normatively rich, the Recommendation lacks direct enforcement tools, its impact depends on domestic uptake and funding.

2.1.3 The EU AI Act

The EU AI Act, which entered into force in 2024, is the most comprehensive binding regulatory framework for AI to date . It classifies systems into prohibited, high-risk, limited-risk, and minimal-risk categories and prescribes obligations on risk management, data quality, transparency, human oversight, robustness, and cybersecurity, including special provisions for general-purpose AI (GPAI) and foundation models .

Implementation is phased; key obligations for GPAI providers and high-risk systems will only bite between 2025 and 2027. Early commentary highlights uncertainty among firms, delays in technical standards, and calls from industry to pause or slow implementation, suggesting challenges ahead for effectiveness [7][23].

2.1.4 NIST AI Risk Management Framework

The NIST AI RMF (2023) provides a voluntary framework to help organizations manage AI risks through four functions: Govern, Map, Measure, and Manage. In 2024 NIST issued a Generative AI Profile (AI 600-1), tailoring the RMF to risks posed by large-scale generative models, including hallucinations, data leakage, and content manipulation [8].

The AI RMF is increasingly referenced as a "gold standard" in US federal guidance and private-sector compliance strategies[22]. However, empirical data suggest that many organizations still lack mature processes, and the framework's voluntary nature limits its reach [11][18].

2.1.5 China's AI regulatory regime

China has adopted some of the earliest binding regulations targeting specific classes of algorithms. These include:

- The 2021 Regulation on Recommendation Algorithms in Internet Information Services
- 2022 provisions on "deep synthesis" (synthetic content)
- 2023 Interim Measures for the Management of Generative AI Services, the first comprehensive binding regulation for generative AI [10].

These instruments combine information control (content moderation, censorship) with economic and social governance objectives (worker protection from algorithmic scheduling, price discrimination). Enforcement relies on filing, audits, and platform accountability. Analyses note that China has moved faster than many democracies in formal rule-making but has weaker transparency and rights safeguards [3][10][21].

2.1.6 African Union Continental AI Strategy

The African Union's Continental AI Strategy, adopted in 2024, is one of the first regional strategies in the Global South . It emphasizes an "Africa-centric, development-focused" approach, with pillars on harnessing AI benefits, building capabilities, minimizing risks, stimulating investment, and fostering cooperation.Implementation will rely on national data protection laws, emerging AI policies, and significant capacity-building. Current analyses highlight limited regulatory resources, infrastructure gaps, and dependency on foreign platforms as persistent constraints [3][9].

2.2 Conceptualizations of AI governance and responsible AI

Recent scoping reviews and systematic analyses reveal a rapidly expanding but fragmented literature on AI governance and responsible AI. Identifying diverse governance practices across the AI lifecycle but find little consensus on how high-level principles are operationalized in design, deployment, and evaluation [1]. Synthesizing 28 AI governance solutions, frameworks, tools, policies and concluding that challenges cluster around unclear responsibilities, lack of enforcement mechanisms, and limited empirical evaluation of effectiveness [2]. Zaidan (2024) emphasizes AI governance as a multi-actor, multi-level problem in a rapidly changing environment, where states, firms, civil-society organizations, and international bodies interact under conditions of regulatory fragmentation and geopolitical competition [3]. A separate line of work proposes unified theoretical frameworks for AI governance that combine legal, ethical, and technical layers, but these remain mostly conceptual and untested at scale [15].

2.3 Implementation and effectiveness of AI governance

The literature documents substantial gaps between governance designs and implementation outcomes:

- **Governmental implementation:** OECD's 2023 report on the state of implementation of the AI Principles finds that while many countries have adopted national AI strategies and created oversight structures, only a minority have robust mechanisms for monitoring impacts or auditing high-risk systems [4][20]. UN e-government and AI-readiness surveys similarly highlight disparities in infrastructure, skills, and regulatory capacity across regions [21].
- **Organizational implementation:** The 2024 AI Index's Global State of Responsible AI survey shows that most large organizations recognize risks like privacy, security, and reliability, but few have fully operationalized comprehensive mitigations; for example, less than 0.6% reported fully implementing all six key data-governance measures [11]. PwC's 2025 Responsible AI survey and IDC/Microsoft surveys similarly report that while over 90% of firms use AI, a majority struggle to move from policies to embedded practices, citing lack of skills, unclear ownership, and inadequate tooling [18][25].
- **Sectoral case studies:** Healthcare a high-stakes domain illustrates the difficulty of turning governance principles into practice. Kim (2025) and Freeman et al. (2025) show that hospitals recognize the need for AI governance structures but face regulatory complexity, fragmented data systems, and limited AI literacy [19][24]. Scoping reviews by Hassan (2024), Nair (2024) and Bouderhem (2024) identify human, organizational, and technical barriers including trust deficits, concerns about autonomy, interoperability issues, and equity risks [19][21][28].

Finally, the OECD's 2025 work on Governing with AI highlights implementation challenges in public administrations, including skills gaps, legacy IT, and fragmented accountability for AI deployments [22].

2.4 Public perceptions, trust, and legitimacy

Public sentiment around AI is increasingly well-documented:

- Pew Research finds that a majority of Americans are more concerned than excited about AI (52% in 2023), with concern clustering around job loss, surveillance, and fairness [13]. A 2025 report comparing AI experts and the public reveals that experts are more optimistic but both groups want stronger oversight and personal control over AI [12][16][22].

- A Brookings nationwide survey and Gallup data show that nearly all Americans now use AI-enabled products weekly, often without recognizing them as AI, while simultaneously expressing strong worries about misinformation, privacy, and labor impacts [14][19].
- A global survey across 47 countries by the University of Melbourne and KPMG reports that two-thirds of respondents use AI regularly and 83% expect benefits, yet 58% consider AI untrustworthy, with higher trust in emerging economies than in advanced ones [15].

Experimental research on automated decision-making and political persuasion underscores that legitimacy and trust depend not only on outcomes but on perceptions of fairness, transparency, and agency. Recent work finds that chatbots can significantly sway political opinions while often being inaccurate, raising new concerns around democratic integrity [24].

These findings indicate that public perceptions are complex, ambivalent, and context-dependent. Legitimacy relies on both substantive protections and a sense that governance frameworks understand and address lived concerns.

2.5 Shadow AI and rogue practices

Emerging work describes shadow AI as the use of AI tools outside organizational approval or governance processes. Surveys in 2024–2025 show:

- Around 59% of US employees use unapproved AI tools at work; among executives this can reach over 90% [16].
- 75% of shadow-AI users admit sharing sensitive data, including customer information, internal documents, and source code .
- In education and other sectors, most staff know colleagues using unauthorized AI, while only a minority report clear institutional policies [16].

These patterns highlight a crucial implementation failure, governance that is misaligned with user needs and perceptions drives practice into the shadows, undermining both risk management and trust.

2.6 Research gap

Existing scholarship richly describes individual frameworks, sector-specific governance, and public attitudes. However, there is less work that **systematically compares major governance**

frameworks across regions, evaluates their implementation effectiveness using empirical indicators, and explicitly analyzes their alignment with public fears and expectations.

This thesis addresses that gap by integrating policy analysis, implementation evidence, and public-perception data into a single comparative framework.

Chapter 3: Methodology

3.1 Research design

This thesis adopts a comparative, mixed-methods synthesis, combining:

1. **Document and policy analysis** of major governance frameworks (EU AI Act, NIST AI RMF, OECD Principles, UNESCO Recommendation, China's regulations, AU AI Strategy).
2. **Secondary empirical analysis** of large-scale surveys and reports on responsible AI adoption, implementation challenges, and public perceptions.
3. **Analytical alignment framework** to assess how governance priorities overlap or diverge from public concerns and how this shapes effectiveness.

The goal is not to generate new primary survey data, but to **integrate and interpret** recent high-quality empirical work to test the research hypothesis.

3.2 Framework selection and scope

The analysis focuses on frameworks that are:

- Widely recognized and influential;
- Either binding (law/regulation) or soft-law with significant normative weight;
- Active and relevant between 2022 and 2025.

The core set includes:

- **Global / soft-law:** OECD AI Principles [4][6]; UNESCO Recommendation on the Ethics of AI [5]; UN Secretary-General's "Governing AI for Humanity" report [25].
- **Regional / binding:** EU AI Act [7]; African Union Continental AI Strategy [9].

- **National / regulatory:** China's recommendation algorithm rules, deep synthesis rules, and generative AI measures [10]; US executive and agency-centered approach, represented via NIST AI RMF [8] and federal guidance building on it.

While other frameworks exist (e.g G7 Hiroshima Process, ISO/IEC 42001), they are used primarily as contextual references.

3.3 Data sources

Secondary data comes from four main categories:

1. **Implementation and governance surveys**
    - OECD implementation reports on AI Principles [4][20]
    - Stanford AI Index 2024 responsible AI chapter and global state of responsible AI survey [11][18]
    - PwC 2025 Responsible AI survey [18][10]
    - IDC/Microsoft and EY surveys on responsible AI and AI-related financial risk [23][25]
    - OECD "Governing with AI" implementation-challenges analysis [22].
2. **Sectoral case studies**
    - Governance and adoption of AI in healthcare institutions [19].
    - Algorithmic accountability in government (e.g., Brazilian auditing systems) [20].
3. **Public opinion and trust surveys**
    - Pew Research Center reports on US public and AI experts [12][13]
    - Global trust and usage surveys (University of Melbourne/KPMG) [15]
    - Brookings and Gallup data on AI use and perceptions [14][19].
4. **Shadow AI and organizational practice**
    - Cybernews, UpGuard, and related surveys on unapproved AI use in workplaces
    - Workday and others on AI distrust and unclear policies [16].

Academic conceptual work on AI governance provides the theoretical lens].

3.4 Analytical framework

The analysis proceeds in two stages.

3.4.1 Evaluating effectiveness

Effectiveness:

1. **Coverage of risk** : whether frameworks address major technical and socio-technical risks (safety, security, fairness, privacy, disinformation, labor impacts, etc.).
2. **Regulatory strength** :whether obligations are binding or voluntary, and what enforcement mechanisms exist.
3. **Institutional capacity** : whether there is evidence of adequate regulatory institutions, resources, and technical capability.
4. **Observed implementation** :survey evidence on organizational adoption of governance measures, sectoral case studies, and incident/impact indicators where available.

This yields a qualitative comparative judgment (high, moderate, low) for each framework-region pair.

3.4.2 Assessing governance–public alignment

Alignment is defined as the degree to which:

- The **risks and values** highlighted by governance frameworks match those most salient to publics
- The **remedies and mechanisms** (e.g., transparency, oversight) correspond to what publics perceive as meaningful protection
- Publics see governance as **legitimate** (inclusive, participatory, responsive).

Alignment is assessed using:

- Content analysis of framework texts (what risks, values, and stakeholders are foregrounded);
- Public-opinion data (what concerns are most salient; attitudes toward regulation)
- Case evidence where public pushback or acceptance has shaped governance (e.g., controversies around surveillance, political persuasion, or content moderation).

3.5 Limitations

This study has several limitations:

- It relies on **secondary data**, which means measurement choices and sampling biases of original studies are inherited.
- It does not provide a quantitative "scorecard" but rather a **structured qualitative comparison**, guided by evidence.
- Public perceptions are often national or regional; findings cannot be assumed to generalize globally.
- The analysis is temporally bounded; a rapidly evolving regulatory landscape may make some conclusions contingent.

Despite these constraints, the synthesis offers a grounded, cross-cutting view that is currently missing in the literature.

Chapter 4 :Results and Analysis

4.1 Effectiveness of major AI governance frameworks

4.1.1 OECD AI Principles and UNESCO Recommendation

Coverage of risk. Both frameworks articulate broad sets of ethical principles: human rights, fairness, privacy, transparency, accountability, robustness, and sustainability [4][5]. They recognize systemic risks such as discrimination, surveillance, and environmental impact. However, they remain **high-level and technology-agnostic**, with limited guidance on emerging capabilities like frontier models or autonomous agents.

Regulatory strength and capacity. These instruments are **non-binding**; they depend on domestic uptake and integration into law. OECD implementation reviews show wide variation in how countries translate principles into enforceable rules [4]. UNESCO's own updates highlight that many member states are still at early stages of implementation [5][25].

Observed implementation. States have launched numerous AI strategies and ethics guidelines, but there is little evidence that these alone have reduced incidents of AI-related harm. OECD and Stanford AI Index data show that despite rapid growth in AI-related regulation, incidents and risk reports continue to rise [11].

Assessment. OECD and UNESCO frameworks have **high normative coverage but low direct effectiveness**, serving as templates that need concrete instantiation.

4.1.2 EU AI Act

Coverage of risk. The Act comprehensively addresses many classes of risk: safety, fundamental rights, transparency, data quality, human oversight, and cybersecurity . Provisions for GPAI and foundation models begin to tackle systemic and cross-sector risks.

Regulatory strength. Obligations are binding, backed by substantial fines (up to 7% of global turnover) and enforced by national authorities and an EU AI Office [23].

Institutional capacity. The EU has relatively strong regulatory institutions, but implementation requires new technical standards and oversight expertise. Delays in codes of practice and standardization, and calls from industry to pause implementation, reveal capacity and clarity issues.

Observed implementation. Since key deadlines fall between 2025 and 2027, empirical evaluation is preliminary. Industry readiness surveys report confusion, resource constraints, and concerns about competitiveness, suggesting that **compliance will be uneven**, especially among SMEs.

Assessment. The EU AI Act has **high potential effectiveness** due to its binding nature and scope, but its actual impact will depend on timely standards, regulator capacity, and the balance between enforcement and innovation.

4.1.3 NIST AI RMF and US patchwork governance

Coverage of risk. The AI RMF and Generative AI Profile cover a broad set of risk dimensions (safety, security, robustness, fairness, privacy, transparency) and provide concrete organizational processes [8]. However, these tools rely on uptake by agencies and firms.

Regulatory strength. The RMF is voluntary, but federal guidance (OMB memoranda, sectoral regulators) increasingly reference it as a benchmark, effectively turning it into a **de facto standard** in some domains [8]. Enforcement remains fragmented across existing laws (consumer protection, anti-discrimination, sector regulation).

Institutional capacity and implementation. Surveys show that many US organizations have AI policies, but only a minority have mature governance practices, responsible AI remains concentrated among “high performers” [11]. EY and Experian data highlight that most

companies experience risk-related financial loss and struggle to implement responsible AI effectively [18][23].

Assessment. The US approach provides flexible but uneven governance, leading firms and regulators can implement strong practices, but there is no systematic, nationwide requirement comparable to the EU AI Act.

4.1.4 China's algorithm and AI regulations

Coverage of risk. Chinese regulations target specific risk classes: harmful or destabilizing content, deepfakes, excessive price discrimination, and worker exploitation under algorithmic management [10][18]. They also impose obligations on generative AI providers regarding training data, content labeling, security assessments, and registration [21][24].

Regulatory strength and capacity. These measures are binding and enforced by the Cyberspace Administration of China and other agencies. Enforcement tends to be strong where political priorities (information control, social stability) are implicated.

Observed implementation. China has produced large numbers of filed algorithms and generative AI services under these rules, and leads in adoption of some controversial use cases, such as continuous automated monitoring, raising global concerns about privacy and accountability [21].

Assessment. China's framework is effective at consolidating state control and setting obligations for providers, but less aligned with global human-rights-based conceptions of trustworthy AI. Its model illustrates that "effectiveness" can be politically contingent.

4.1.5 African Union Continental AI Strategy

Coverage of risk. The AU strategy explicitly addresses development, inequality, and capacity-building alongside ethical and security risks [9][10]. It recognizes both opportunities and risks, including bias, exclusion, and dependency on foreign technology.

Regulatory strength and capacity. The strategy is not, itself, binding law; it provides a roadmap and calls for harmonized regulatory approaches. Analyses highlight significant gaps in infrastructure, skills, and resources for implementation across the continent [3][9][21].

Observed implementation. A small but growing number of African states are adopting AI strategies or embedding AI in data-protection regimes. However, there is limited evidence yet of robust enforcement of AI-specific obligations.

Assessment. The AU framework has high relevance but currently low implementation capacity, risking a widening gap between normative aspirations and actual protections.

4.2 Cross-cutting patterns

Three broad patterns emerge.

4.2.1 Proliferation of rules vs. lagging implementation

Stanford's AI Index documents rapid growth of AI-related laws and regulations globally, alongside a rise in documented AI incidents and harms [11][17]. OECD and UN reports similarly show more strategies than mature oversight structures [4][21][22].This indicates that formal rule-making has not yet translated into commensurate risk reduction.

4.2.2 Voluntary vs. binding regimes

Voluntary frameworks (OECD Principles, UNESCO Recommendation, NIST RMF) have high flexibility and broad uptake but limited authority over high-risk or non-cooperative actors. Binding laws (EU AI Act, Chinese measures) carry stronger enforcement but can face resistance, calls for delays, and concerns about innovation [7][10][23]. The overall landscape is a patchwork, with cross-border systems subject to multiple, sometimes conflicting, regimes.

4.2.3 Sectoral imbalance

High-stakes domains like healthcare, welfare, and criminal justice often lag in governance maturity despite high potential harms. Healthcare studies show that governance structures are emerging but face regulatory complexity, data fragmentation, and low AI literacy [19][21]. By contrast, financial and large tech firms often deploy more sophisticated risk management, motivated by regulatory pressure and brand concerns [11][18].

4.3 Governance–public perception alignment

4.3.1 What governance frameworks prioritize

Across frameworks, recurring priorities include:

- Fairness / non-discrimination;
- Transparency and explainability;
- Safety and robustness;
- Privacy and data protection;
- Accountability and human oversight.

These align broadly with academic discourse on trustworthy AI [1][2][3][11].

However, **lived public concerns** often extend beyond these categories:

- Fear of **job loss and economic displacement**
- Anxiety about **surveillance and loss of autonomy**
- Worry about **misinformation, political manipulation, and erosion of democratic deliberation**
- Frustration with lack of **participation and control** in AI governance.

Pew's 2023–2025 surveys show that American adults are more concerned than excited about AI; they associate AI with worsening inequality and relationships, and they consistently call for stronger oversight [12][13][22]. Global surveys show higher optimism in emerging economies, but widespread perceptions that AI is untrustworthy and poorly regulated [15][19].

Many frameworks **address technical risks** but say little about economic precarity, political persuasion, or everyday experiences of opacity and dependency.

4.3.2 Shadow AI as a symptom of misalignment

The prevalence of shadow AI demonstrates how misalignment affects behavior. Workers use unapproved tools because they perceive them as useful, accessible, and necessary, while viewing official policies as slow, unclear, or irrelevant [16]. At the same time, organizations fear data leakage, compliance violations, and reputational risks [18][22].

Where governance frameworks do not integrate user-centric design and realistic workflows, they **push practice underground**, weakening both security and accountability.

4.3.3 Democratic legitimacy and persuasive AI

The UK AI Safety Institute's study on chatbots shows that general-purpose models can be highly persuasive on political topics while often being inaccurate [24]. This risk AI's role in shaping political opinions and public discourse is only partially addressed in current governance regimes, which focus more on safety, bias, and basic transparency than on information power and epistemic harms. Publics express concern about misinformation and manipulation, yet most frameworks have not fully integrated mechanisms such as content provenance, political-communication safeguards, or limits on automated persuasion.

4.4 Consequences: Trust, compliance, and effectiveness

The misalignment between governance and public perceptions leads to several concrete outcomes:

1. **Erosion of trust.** Even as frameworks multiply, surveys show enduring skepticism that AI will benefit ordinary people or be deployed responsibly [12][13][15][19].
2. **Partial compliance.** Organizations implement parts of responsible AI practices, often those easiest to formalize (policies, committees), while more demanding practices (comprehensive monitoring, fairness audits, incident reporting) lag [11][18].
3. **Rogue and shadow practices.** Workers and departments adopt AI tools outside official channels, undermining centralized governance and increasing unmonitored risk [16].
4. **Inequitable global protection.** Regions with strong regulatory capacity (EU, parts of North America) can enforce protections; others (many African and low-income states) rely on soft-law and imported tools, creating asymmetric exposure to risks [3][9][21][22].

Taken together, these findings support the hypothesis that frameworks are structurally insufficient in their current form, not primarily because of conceptual flaws, but because of reliance on high-level principles, voluntary adoption, and limited integration of public experience and capacity realities.

Chapter 5:Discussion and Recommendations

5.1 Interpreting the findings against the hypothesis

The evidence assembled here broadly supports the central hypothesis:

- **Structural insufficiency:** Frameworks lean heavily on broad principles, risk-taxonomies, and voluntary adoption, with limited binding obligations outside the EU and China. Even in binding regimes, enforcement is nascent and resource-constrained.
- **Misalignment with public fears:** Governance focuses on technical properties (robustness, bias, privacy) and organizational process, while publics emphasize economic security, autonomy, surveillance, and manipulation.
- **Consequences for trust and effectiveness:** Misalignment and weak enforcement correlate with shadow AI, fragmented compliance, and persistent distrust, which in turn undermine governance aims.

However, the picture is nuanced. The EU AI Act and China's generative AI measures demonstrate that stronger, binding regimes can be implemented, albeit with trade-offs (innovation concerns, rights issues). The AU strategy shows how development-oriented governance can foreground equity and sovereignty, even if implementation capacity lags. Thus, the problem is not simply that frameworks exist; it is that they are not yet designed and implemented as socio-technical systems that reflect how people actually use, experience, and contest AI.

5.2 Design principles for more effective, public-aligned governance

Based on the analysis, seven design principles emerge.

5.2.1 From AI to intelligent systems

Current frameworks often target AI as a category, but many high-impact systems algorithmic decision-support, scoring, automated monitoring are not branded as AI. Governance should apply to intelligent systems defined by function, autonomy, and systemic impact, not by technical label.

This would:

- Bring algorithmic management, scoring, and surveillance systems squarely under governance
- Avoid loopholes where risky systems evade oversight because they do not meet narrow AI definitions.

5.2.2 Hardening soft-law through enforceable mechanisms

Soft-law instruments (OECD, UNESCO, NIST) should be coupled with:

- **Binding requirements** for high-risk domains (e.g health, welfare, policing, critical infrastructure)
- **Third-party audits and certifications** (e.g ISO 42001-type AIMS) recognized by regulators
- **Incident reporting obligations**, akin to safety reporting in aviation.

These mechanisms move governance from aspirational to operative, particularly for actors that are not voluntarily inclined to adopt strong practices.

5.2.3 Institutionalizing participation and co-design

Governance that is perceived as legitimate must be participatory:

- National frameworks should require **stakeholder consultations**, including workers, affected communities, and civil-society organizations, for high-risk deployments.

- Citizen assemblies or deliberative mini-publics could be convened to guide policies on politically sensitive uses such as surveillance, predictive policing, and political communication.

- At organizational level, AI governance should include **worker councils or ethics boards with representation from those affected**.

Empirical research shows that legitimacy perceptions hinge on process fairness and inclusiveness as much as on outcomes [16][20].

5.2.4 Building capacity and shared infrastructure, especially in the Global South

To avoid governance becoming a luxury of wealthy states:

- Global initiatives (UN, OECD, AU, UNESCO) should prioritize **funding regional evaluation labs, shared testbeds, and open-source tools** for risk assessment and auditing.
- Capacity-building should focus not only on technical skills, but on **regulatory craft**, socio-technical analysis, and community engagement [9][21][22][25].
- International financial institutions could condition AI-related investments on adherence to baseline governance standards.

Without such investments, soft-law norms will remain aspirational in many contexts.

5.2.5 Designing for shadow AI rather than ignoring it

Given the persistence of shadow AI:

- Organizations should be required (or strongly encouraged) to conduct AI use audits, not just system inventories.
- Governance frameworks should support safe AI sandboxes where employees can experiment with approved tools under protective controls (e.g., data redaction, DLP) rather than prohibiting AI outright.
- Training and communication must address why rules exist, what risks matter, and how employees can benefit from AI safely.

This reframes governance as enabling safe use rather than simply restricting tools, reducing incentives to go underground [16][18].

5.2.6 Addressing epistemic and political harms explicitly

Frameworks need to confront informational and democratic risks head-on, including:

- Standards for transparency and provenance in political ads and automated messaging
- Restrictions or heightened scrutiny for AI systems used to target political persuasion
- Support for public-interest research on AI's impact on discourse and democracy [24].

Without explicit treatment of these harms, frameworks risk being seen as technocratic and blind to core public concerns.

5.2.7 Embedding Developer and Organizational Accountability Through Personal and Corporate Liability

Current AI governance frameworks overwhelmingly target systems, not the people or institutions that design, deploy, or misuse them. Across the EU AI Act, UNESCO Recommendations, OECD Principles, and NIST AI RMF, accountability is conceptualized mainly as documentation, transparency, or organizational risk management. These mechanisms rarely attach direct consequences to developers, executives, or decision-makers who release unsafe or deceptive intelligent systems. As a result, harms are often treated as technical failures rather than failures of judgment, governance, or corporate responsibility.

A growing literature in technology governance argues that the absence of personal and corporate liability creates moral hazard: organizations can externalize harms onto the public while reaping the benefits of deploying rapidly scaled intelligent systems. In many sectors, the most damaging AI systems are not "accidents," but foreseeable risks ignored due to competitive pressure, inadequate testing, or intentional misuse. Existing frameworks rarely impose professional accountability (e.g., certification requirements), criminal liability for reckless deployment, or civil penalties for negligent system release.

To address this gap, a more effective governance model must extend beyond system-focused controls and include:

1. Personal liability for executives and lead engineers who knowingly release systems that violate safety, discrimination, or transparency obligations
2. Corporate fines proportionate to global revenue, similar to GDPR, for negligent deployment or harmful misuse
3. Mandatory disclosure of high-risk models, with penalties for concealment
4. Regulatory authority to ban repeat-offending developers or vendors from deploying AI in certain sectors.

These mechanisms shift the focus from AI as a rogue actor to the humans and institutions who design and profit from it, reinforcing the principle that technological harm is not inevitable but preventable. By attaching real-world consequences to irresponsible behavior, governance can move from symbolic principles to substantive deterrence, enhancing both effectiveness and public trust.

5.3 Implications for policymakers and practitioners

For policymakers, the analysis suggests:

- Investing as much in implementation infrastructure (institutions, skills, audits) as in drafting new rules
- Using hybrid models (binding rules + codes of practice + soft-law) to balance legal certainty and flexibility
- Integrating public opinion monitoring into governance cycles, treating trust metrics as policy indicators.

For organizations, it implies:

- Elevating AI governance from a compliance exercise to core risk and strategy function
- Using frameworks like NIST AI RMF and ISO 42001 not only for certification, but for continuous improvement and transparency
- Engaging workers and users in the design of AI-enabled workflows to align governance with practice.

For researchers, there is a need for:

- More empirical evaluation of governance effectiveness, including natural experiments where new laws come into force [1][2][11][26]
- Comparative studies of citizen and stakeholder perceptions across different governance regimes
- Development of metrics for alignment between governance and public sentiment, beyond simple trust scores.

Conclusion

Although the existing global, regional, and national AI governance frameworks provide an essential foundation for guiding the development and deployment of intelligent systems, the evidence presented in this thesis demonstrates that they remain only partially effective without complementary mechanisms that address real-world implementation gaps, public concerns, and emerging forms of technological risk. Far from suggesting that these frameworks be discarded,

the findings show that they can serve as strong structural anchors if reinforced with more robust enforcement tools, clearer accountability pathways, participatory oversight, and capacity-building initiatives, especially in under-resourced regions. Data across organizations, sectors, and jurisdictions reveal persistent loopholes, uneven adoption, and misalignment between what governance systems prioritize and what the public fears or expects, leading to shallow compliance, shadow AI practices, and declining trust. Strengthening these frameworks therefore requires not only technical refinement but a deeper commitment to protecting people by ensuring that developers, deployers, and decision-makers are held responsible for harmful or negligent actions and by creating governance processes that are transparent, anticipatory, and responsive to societal realities. In this sense, the conclusion is not that we need an entirely new governance paradigm, but that we must build more vigilant, accountable, and empirically grounded layers on top of the systems already in place, closing the gaps that currently allow intelligent systems to outpace the protections intended to safeguard human dignity and collective well-being.

Going forward, policymakers should strengthen existing AI governance frameworks by embedding enforceable accountability measures, expanding regulatory capacity, and ensuring that oversight mechanisms reflect the concerns and lived experiences of the public. This means reinforcing current standards with clearer liability pathways for harmful development and deployment practices, investing in independent evaluation and auditing infrastructure, and prioritizing participatory processes that give communities a direct role in shaping how intelligent systems are governed. By closing the loopholes identified in the evidence and building more responsive, transparent, and people-centered safeguards on top of current frameworks, governments can create a governance ecosystem that not only manages risk more effectively but also restores public trust and ensures that intelligent systems are developed and deployed in ways that genuinely advance societal well-being.

References

1 Papagiannidis, E., Marikyan, D., Rana, N.P., 2024. Responsible artificial intelligence governance: A review. *Technological Forecasting and Social Change*, 198, 122840. https://doi.org/10.1016/j.jsis.2024.101885

2 Batool, A., Alam, M., Ahmed, S., 2025. AI governance: A systematic literature review. *AI and Ethics*, 5(1), pp. 33–58. https://doi.org/10.1007/s43681-024-00653-w

3 Zaidan, E., 2024. AI governance in a complex and rapidly changing world. *Humanities and Social Sciences Communications*, 11, 105. https://doi.org/10.1057/s41599-024-03560-x

4 OECD, 2023. *The State of Implementation of the OECD AI Principles Four Years On.* OECD AI Papers No. 3. OECD Publishing. https://doi.org/10.1787/835641c9-en

5 UNESCO, 2021. *Recommendation on the Ethics of Artificial Intelligence.* United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000381137

6 OECD, 2023. *OECD Framework for the Implementation of Trustworthy AI.* OECD Publishing. https://oecd.ai/en/ai-principles

7 European Commission, 2024. *Artificial Intelligence Act (EU AI Act).* Official Journal of the European Union. https://artificialintelligenceact.eu › the-act

8 NIST, 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

9 African Union Commission, 2024. *Continental Strategy for Artificial Intelligence.* AU Digital Transformation Agenda.https://au.int/sites/default/files/documents/44004-doc-EN-_Continental_AI_Strategy_July_2024.pdf

10 Cyberspace Administration of China (CAC), 2021. *Provisions on Algorithmic Recommendation in Internet Information Services.*

CAC, 2022. *Administrative Provisions on Deep Synthesis Internet Information Services.*

CAC, 2023. *Interim Measures for the Management of Generative Artificial Intelligence Services.*

11 Stanford Institute for Human-Centered Artificial Intelligence, 2024. *AI Index Report 2024.* Stanford University. https://hai.stanford.edu/ai-index/2024-ai-index-report

12 Pew Research Center, 2023. *Public Attitudes Toward Artificial Intelligence in the United States.* *https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/*

13 Pew Research Center, 2024. *How AI Experts and the Public View Artificial Intelligence.* https://share.google/XOG3XlQXptiRHSejU

14 Brookings Institution, 2024. *AI Perceptions, Concerns, and Use Patterns in the American Public.* Brookings Governance Studies.

15 KPMG Australia, University of Melbourne, 2024. *Global AI Trust and Adoption Index.*

16 Cybernews Research Group, 2024. *Shadow AI in the Workplace: Employee Use of Unapproved AI Tools.*

17 Stanford University, 2024. *AI Incidents and Harms Database: 10-Year Trend Analysis.*

18 PwC, 2025. *2025 Responsible AI and Emerging Technology Survey.* PwC Global Risk Insights.

19 Freeman, J., Kim, S., Patel, R., 2025. Governance readiness and adoption barriers for AI in clinical environments: A multi-hospital study. *Journal of Medical Systems*, 49(2), 55.

20 OECD, 2025. *Governing with AI: Implementation Capacity, Challenges, and Public Sector Applications.* OECD Digital Government Studies.

21 United Nations Department of Economic and Social Affairs, 2024. *E-Government Survey 2024: AI Readiness and Digital Public Infrastructure.*

22 Office of Management and Budget, 2024. *Advancing Responsible Artificial Intelligence in Government.* OMB Memorandum.

23 IDC, Microsoft, 2024. *Responsible AI in Enterprise: Adoption, Maturity, and Financial Risk Assessment.*

24 Department for Science, Innovation and Technology, 2024. Evaluating Political Persuasion Risks of Large Language Models. AI Safety Institute Report.

25 United Nations, 2023. Governing AI for Humanity: Report of the UN Secretary-General's Advisory Body on AI.

26 Department for Science, Innovation and Technology, 2024. *Evaluating Political Persuasion Risks of Large Language Models.* AI Safety Institute Report.

26 Nair, V., 2024. Human, organizational, and technical barriers to safe AI deployment in healthcare: A scoping review. *BMC Health Services Research*, 24, 773.

27 Bouderhem, S., 2024. Equity, interoperability, and trust in AI-enabled healthcare systems: A systematic analysis. *Health Policy and Technology*, 13(2), 100821.

28 Hassan, R., 2024. AI governance in healthcare: Autonomy, trust, and system-level challenges. *International Journal of Medical Informatics*, 189, 105303.