Toward a Human-Governed Intelligence Infrastructure

AI Governance for National Security Systems in High-Consequence Environments

Matthew McClendon
National Security Policy Advisor — Emerging Infrastructure
Department of Homeland Security — Office of Intelligence & Analysis

FOREWORD

The United States intelligence apparatus was built for speed, reach, and asymmetry. It was not built for mutual visibility between citizen and state, nor for systems that must interpret human behavior in real time, at scale, and under conditions of ambiguity.

That constraint is no longer theoretical.

Artificial intelligence now sits inside the intelligence lifecycle, shaping how signals are surfaced, interpreted, and acted upon. At the same time, the nature of threat has shifted. Risks are no longer exclusively external or state-bound. They emerge within domestic systems, through digital environments, and across populations whose relationship to government is often defined by mistrust.

This creates a structural tension.

Systems optimized for detection now operate in environments where misinterpretation carries strategic cost. False positives erode civil legitimacy. Opaque systems degrade trust. Misread human context produces flawed intelligence, even when the underlying data is correct.

This paper advances a position:

AI-enabled intelligence systems must be governed as high-consequence infrastructure.
They must remain human-directed, auditable, and accountable to the populations they affect.

This is not a departure from intelligence practice. It is an adaptation to the conditions under which intelligence now operates.

EXECUTIVE SUMMARY

The integration of AI into national security systems introduces a new class of risk: not only the failure to detect threats, but the misinterpretation of human behavior at scale.

Traditional intelligence frameworks were not designed to manage:

- probabilistic outputs presented as actionable insight
- opaque model reasoning
- feedback loops between institutional action and public perception
- the downstream societal impact of false positives

This paper proposes a human-governed intelligence infrastructure, defined by:

- enforced human oversight across all AI-assisted decisions
- auditability of model outputs and decision pathways
- explicit management of false positive harm as a national security risk
- integration of contextual, cultural, and behavioral intelligence into analysis
- governance structures that enable contestability and correction

The objective is clear:

To ensure that AI strengthens intelligence outcomes without degrading public trust, civil legitimacy, or operational accuracy.

I. BACKGROUND: FROM SECRECY TO COMPLEXITY

Modern intelligence systems were designed in a context defined by:

- state-based adversaries
- centralized information control
- clear boundaries between domestic and foreign threat

That context no longer holds.

Today's threat environment is:

- decentralized and networked
- shaped by open-source and participatory information flows
- influenced by social, behavioral, and economic conditions
- tightly coupled with domestic institutional trust

This shift introduces a new failure mode.

The challenge is no longer access to data.
The challenge is interpretation under conditions of ambiguity and scale.

AI accelerates this dynamic. It increases the speed of signal detection while introducing opacity into how those signals are derived. Without governance, this creates systems that are efficient, but not necessarily accurate, and scalable, but not necessarily legitimate.

II. WORKING DEFINITION: HUMAN-GOVERNED INTELLIGENCE

A human-governed intelligence infrastructure is one in which:
AI systems inform, but do not conclude

- human operators retain decision authority in all high-consequence contexts
- model outputs are interpretable, auditable, and contestable
- contextual and lived experience are treated as valid inputs to analysis
- feedback loops exist between institutional action and public impact

This model reframes intelligence as a system of interpreted signals, not raw detection.

It recognizes that the end user of intelligence is not only the policymaker or agency, but the public itself, whose behavior, trust, and perception directly shape national security outcomes.

III. THE PROBLEM: CIVICALLY ILLEGIBLE INTELLIGENCE

An intelligence system can be technically correct and still be operationally wrong if it is civically illegible.

Common failure patterns include:

- reduction of complex human behavior into static risk categories
- reliance on historically biased datasets without corrective mechanisms
- interpretation of distrust or instability without examining root causes
- separation of intelligence analysis from the civil systems that generate observed signals

Example:

An individual experiencing a mental health crisis may be surfaced as a threat indicator within surveillance systems. The systemic failures that produced that condition remain unexamined. The resulting intelligence product reflects observable behavior, but not causal reality.

This creates distortion.

Over time, repeated distortions degrade both decision quality and institutional credibility.

IV. AI GOVERNANCE IN HIGH-CONSEQUENCE ENVIRONMENTS

Governance in this context is not advisory. It is binding.

AI systems operating within national security contexts must be governed as critical infrastructure.

1. AI as Advisory System, Not Decision Authority

   AI systems must operate within clearly defined control modes:
   - Human-in-the-loop: required for all domestic, rights-impacting decisions
   - Human-on-the-loop: permitted in bounded, lower-risk monitoring contexts
   - Fully autonomous: prohibited in domestic intelligence applications

   No AI-generated output should trigger action without human review where civil impact is present.

2. Visibility of Uncertainty

   All AI outputs must include:
   - confidence levels
   - contributing factors or feature visibility
   - known limitations or blind spots

   Confidence is not truth.
   Uncertainty must be visible at the point of decision.

3. False Positives as Strategic Risk

   False positives are not operational noise. They are systemic risk.

   They result in:
   - wrongful targeting or escalation
   - erosion of public trust
   - degradation of future signal quality as communities disengage

Governance frameworks must explicitly track, audit, and minimize false positive harm alongside traditional detection metrics.

4. Auditability and the Chain of Interpretation

   Every AI-assisted intelligence action must be reconstructable.

   This requires a Chain of Interpretation, including:
   - source data inputs
   - model outputs and transformations
   - human analyst decisions
   - final operational actions

   This record enables oversight, accountability, and post hoc review under legal and policy scrutiny.

5. Contestability and Feedback

   Communities and individuals affected by intelligence outputs must have mechanisms to:
   - contest interpretations
   - provide contextual correction
   - surface misclassification or harm

   Without contestability, systems become self-reinforcing and degrade over time.

6. Red Teaming with System-Impacted Populations

   Adversarial testing must extend beyond technical scenarios.

   It must include:
   - individuals and communities historically subject to surveillance
   - simulations of bias, misclassification, and overreach
   - stress testing of systems under real-world ambiguity

This approach reveals failure modes invisible to purely technical review.

7. Prohibited Design Patterns

   The following system characteristics should be explicitly disallowed:

   - autonomous domestic threat classification without human oversight
   - opaque risk scoring systems without explanation
   - models that conflate correlation with intent
   - training pipelines that rely on biased enforcement data without mitigation

   Governance requires not only what is built, but what is prevented.

   Decision authority must remain human, not as a safeguard of last resort, but as a continuous control layer across the intelligence lifecycle.

V. INSTITUTIONAL FLUENCY IN HUMAN CONTEXT

Intelligence agencies must expand capability beyond traditional analytic domains.

Required areas of fluency include:

- behavioral and mental health dynamics
- structural inequality and systemic risk factors
- data ethics and algorithmic accountability
- narrative and sentiment analysis across populations

This is not a soft capability. It is required for accurate interpretation.

Without contextual fluency, increased data volume produces increased misread.

VI. ARCHITECTING TRUST AS INFRASTRUCTURE

Trust is not a communication strategy. It is a structural outcome.

A human-governed intelligence infrastructure must include:

- Transparency nodes: controlled release of non-classified data for external interpretation
- Human Impact Review Boards: formal evaluation of downstream harm, with escalation authority tied to operational review thresholds and the ability to trigger mandatory reassessment of intelligence outputs
- Feedback channels: mechanisms for contesting and correcting institutional outputs
- Training reform: integration of ethics, power literacy, and system design into analyst development

These structural changes introduce tension within existing operational frameworks. Governance reform requires not only system redesign, but institutional alignment around accountability, authority, and acceptable risk. Without that alignment, even well-designed systems will fail under pressure.

Trust emerges when systems are inspectable, accountable, and responsive.

VII. APPLICATION TO MODERN THREAT ENVIRONMENTS

This governance model applies directly to:

- domestic extremism detection
- disinformation and narrative warfare
- border and migration intelligence
- financial and cyber-enabled threat networks
- public unrest and protest analysis

In each domain, the primary failure mode is not lack of data.

It is the misinterpretation of human context.

AI increases the scale of this risk. Governance determines whether that scale produces clarity or distortion.

VIII. FAILURE MODES OF NON-GOVERNED SYSTEMS

Without human-centered governance, AI-enabled intelligence systems will:

- amplify biased historical patterns
- increase false positive rates under the appearance of precision
- produce decisions that cannot be explained or defended
- degrade public trust, reducing future signal availability
- create feedback loops where institutional action reinforces flawed models

These failures do not remain technical.

They become societal.

IX. STRATEGIC OUTCOMES

A human-governed intelligence infrastructure produces:

- improved accuracy through contextual interpretation
- reduced false positives and associated harm
- increased institutional legitimacy and public trust
- stronger signal quality over time
- greater resilience under conditions of uncertainty

This is not a tradeoff between ethics and effectiveness.

It is an alignment of the two.

CONCLUSION: GOVERNANCE DEFINES CAPABILITY

AI will not replace intelligence analysts.
It will amplify their influence.

Governance determines whether that amplification produces clarity or distortion.

The future of national security intelligence is not defined by how much data we can process, but by how well we can interpret human reality under pressure.

Systems that remain opaque, unaccountable, or detached from lived experience will fail, regardless of technical sophistication.

A human-governed intelligence infrastructure does not reduce capability.

It ensures that capability remains aligned with the society it is meant to protect.

This is not theoretical.

This is not a theoretical framework. It is the minimum condition for operational readiness in an AI-mediated intelligence environment.