Algorithmic Surveillance and Population Control of Palestinians

A Game-Theoretic Analysis of Israel's Blue Wolf System in Hebron

Giorgio Coppola and Giulia Maria Petrilli

GRAD-E1489: Algorithmic Game Theory and Governance
Hertie School
Fall 2025

Abstract

This paper models algorithmic surveillance as a strategic game of incomplete information, using Israel's Blue Wolf biometric identification program in Hebron as a case study. Building on the preventive repression framework of Dragu and Lupu (2021), we examine how algorithmic classification and risk scoring create strategic dynamics between state surveillance apparatus and targeted populations. The analysis suggests a mechanism through which technological advances in surveillance can be associated with systematically harsher control policies. We introduce an interpretation suggesting that algorithmic errors and false positives may be tolerated or even strategically advantageous under certain political objectives. The paper contributes to debates on algorithmic governance by demonstrating how formal game theory can illuminate the political logic of surveillance systems.

Keywords: Algorithmic surveillance; game theory; incomplete information; Israel–Palestine; biometric identification; preventive repression

1 Introduction

Algorithmic systems aimed at classifying, tracking, and predicting the behavior of populations are deeply debated in democracies, but they are already in place where people have fewer freedoms. When used by authoritarian governments toward populations whose rights are limited, such technology creates strategic environments. Risk scoring, facial recognition, and movement prediction technologies create continuous games between surveillance authorities and those subjected to monitoring. In fact, the latter might choose to adapt their behavior to defy such systems, and in turn, the surveillance authorities may be able to learn from adaptation and change their strategy accordingly. Therefore, this paper asks: How do targets of algorithmic surveillance adapt their behavior? Does this strategic interaction push systems toward harsher policies and/or systematic errors? And crucially, are such errors tolerated or even instrumentalized because they serve broader goals of population control? We address these questions through analysis of Israel's Blue Wolf system, a biometric surveillance program deployed by the Israeli Defense Forces (IDF) in the occupied West Bank, with particular intensity in Hebron. Blue Wolf combines facial recognition technology, an extensive biographical database (Wolf Pack), and a color-coded risk classification system to allow instantaneous identification and sorting of Palestinian civilians (Dwoskin, 2021; Amnesty International, 2023). Israeli authorities have framed these technologies as fundamental for both enabling a "frictionless" security policy, reducing invasive checkpoints while maintaining security control, and for facilitating the occupation, defining the system as "the Facebook of Palestinians" and Hebron as a "Smart City," highlighting their dual logic of surveillance (Dwoskin, 2021). According to experts, Israel's use of facial recognition is often described as among the most advanced deployments of such technology by a country seeking to control a subject population (Fatafta and Nashif, 2017). Ethnographic research reveals that Blue Wolf intensifies rather than reduces the violence of occupation, eroding Palestinian social life and private space (Goodfriend, 2023). It is important to note that, even if Israel is commonly classified as a democracy in terms of electoral institutions, a substantial literature argues that its governance toward Palestinians is characterized by stratified citizenship and coercive control practices consistent with illiberal or authoritarian logics (Dayan, 2022; Gidron, 2023; Jamal, 2007; Smooha, 2002; Rouhana, 1997; Peled, 2008).

This work develops a game-theoretic framework to analyze the strategic logic underlying such systems. Building on Dragu and Lupu (2021)'s model of digital authoritarianism and preventive repression, we extend their analysis to incorporate algorithmic classification as probabilistic identification, systematic errors, and behavioral adaptation by surveilled populations. In our static setting, the model highlights an escalation mechanism: improvements in surveillance capacity raise the incentives to monitor, while simultaneously inducing costly adaptation by the targeted population, sustaining pressure toward expanded control even when the prevalence of genuine threats is low. The paper proceeds as follows. Section 2 situates the analysis within the relevant literature. Section 3 presents the empirical context of Blue Wolf in Hebron. Section 4 develops the formal model. Section 5 analyzes equilibrium behavior and comparative statics. Section 6 discusses policy implications and limitations. Section 7 identifies a venue for this case to be modeled as a signaling game, providing direction for further research.

2 Theoretical Framework

Dragu and Lupu (2021) provide the foundational framework for this analysis. They model the strategic interaction between an authoritarian government engaging in preventive repression and an opposition group attempting to mobilize dissent. Their key insight is that digital technology has dual effects: it lowers both the barriers to state surveillance and the barriers to opposition mobilization. Within this strategic context, they demonstrate that technological advancement consistently increases equilibrium levels of preventive repression, even when technology also benefits opposition groups. Three results from their analysis are particularly relevant. First, technological innovation unconditionally increases government preventive repression in equilibrium. Second, the probability that government successfully prevents opposition mobilization increases with technological development. Third, authoritarian governments prefer technologies that reduce the cost of preventive control and will strategically allow or block technologies based on their differential effects. We extend this framework in two directions. First, we incorporate incomplete information about individual 'threat types,' i.e., the fact that the state cannot perfectly distinguish between those who pose real security threats and ordinary civilians. This generates classification errors with strategic consequences. Second, we model behavioral adaptation by surveilled populations, i.e., changes in movement, communication, and social behavior in response to algorithmic monitoring.

Critical surveillance studies emphasize that algorithmic systems do not simply observe populations but actively constitute them as objects of governance (Browne, 2015; Benjamin, 2019a). Risk scoring and predictive analytics impose categories that shape subsequent treatment, creating what Harcourt (2007) terms 'actuarial justice.' In contexts of colonial rule, such systems articulate with longer histories of population management and territorial control (Zureik, 2011; Shalhoub-Kevorkian, 2015). Byler (2022)'s concept of 'terror capitalism' illuminates how surveillance technologies serve dual functions: extracting value (valuable information) through data accumulation while facilitating state control over minoritized populations. In the Palestinian context, Goodfriend (2023) documents how Blue Wolf is functional to the simultaneous intensification of control over Palestinians and dispossession of their land and resources, while enabling capital accumulation within Israeli society. Additionally, studies suggest that algorithmic 'errors,' such as misclassifications, may not simply be technical failures to be minimized, but may serve political functions within regimes of population control. Investigative journalism has clarified that in the context of the search and elimination of targets, the Israeli army has dramatically expanded its definition of "human target" since October 7th, when Hamas-led militants launched a deadly assault on southern Israeli communities (Abraham, 2024). From being a term used to designate senior military operatives, under "Operation Iron Swords," the army started to designate all operatives of Hamas's military wing as human targets, irrespective of their rank or military importance. Being now the number of targets in the thousands, the IDF recurred to the use of a new artificial intelligence system, called Lavender, to add people to their 'kill list.' The approval to automatically adopt Lavender's kill lists was only granted about two weeks into the war, even if the program is relatively highly prone to errors and misclassifications (Abraham, 2024). This precedent gives reasons to investigate further whether other systems

used for population control might have similar faults. Additionally, detecting many real security threats would reinforce the counter-terrorism frame that benefits the political goals of Israel and its allies, which suggests that a misclassification might not only be unaccounted for, but even desirable. The literature gives reasons to believe that misclassification is not fully disciplined by institutional incentives and may persist because it may align with broader political interests. The game-theoretic framework developed below formalizes this intuition.

3 Empirical Context: Blue Wolf in Hebron

Blue Wolf is a mobile application deployed by the IDF that allows soldiers to identify Palestinians through facial recognition and biometric matching. The system connects to Wolf Pack, a database containing biographical information, family histories, employment records, and assigned security ratings for Palestinians across the West Bank. The program is paralleled by Red Wolf, which extends this system to automated facial recognition at checkpoints in the West Bank. In the context of Blue Wolf, when a soldier photographs a Palestinian, the app retrieves their profile and displays a color-coded risk assessment: red (detain), yellow (delay for further questioning), or green (release). The system's development relied on intensive data collection. Soldiers were assigned quotas, reportedly fifty photographs per shift, and competed for prizes based on the number of 'pairings' achieved (Kubovich, 2022). This gamification of surveillance (Benjamin, 2019b) transformed military patrols into data harvesting operations, with Palestinian civilians serving as data sources for algorithmic refinement.

Ethnographic research documents how Blue Wolf transforms everyday life in Hebron: movement through the Old City becomes an ongoing encounter with checkpoints, cameras, and soldiers equipped with biometric devices, prompting residents to self-police their mobility and domestic routines. Interviewees describe avoiding monitored thoroughfares and contending with unpredictable stops and scans; at home, surveillance infrastructures installed on rooftops and the possibility of unannounced soldier entry reshape intimate life. Some residents report staying inside to be present when soldiers arrive for maintenance, keeping hijabs and "outside" garments on indoors, and even abandoning patios or other semi-private spaces that now feel exposed to cameras. Attempts to create visual barriers (e.g., hanging tarps) are quickly dismantled, underscoring how adaptation is both pervasive and constrained under algorithmic monitoring (Goodfriend, 2023). Critically, the system's errors generate significant consequences. Such errors, rather than undermining the system's legitimacy, appear absorbed into its operation. The possibility of misclassification may become a mechanism of control, as residents must assume any interaction with soldiers could result in detention regardless of actual behavior.

4 The Model

4.1 Setup

We model the interaction between a surveillance authority (the Israeli military, denoted G) and a population of civilians subject to monitoring (Palestinians in Hebron, denoted P). Following

Dragu and Lupu (2021), this is a simultaneous-move game where G chooses a level of surveillance effort and P chooses a level of behavioral adaptation.

Players and Types. The population P consists of a set of individuals. Each individual i has a private type $\theta_i \in \{T, N\}$, where T denotes 'threat' (e.g., engaged in violent resistance activities) and N denotes 'non-threat' (ordinary civilian, e.g., possibly engaged in social and political activities). Let $\lambda \in (0, 1)$ denote the prior probability that a randomly selected individual is type T. This prior is common knowledge, but individual types are private information.

Government Action. The government chooses surveillance intensity $s \in [0, \bar{s}]$, which determines the probability of detecting and correctly classifying individuals. Higher s increases both the probability of identifying actual threats and the probability of false positives among non-threats. Let:

- $\pi_T(\cdot)$ = probability of correctly identifying a type T individual (true positive rate)
- $\pi_N(\cdot)$ = probability of incorrectly flagging a type N individual (false positive rate)

Both functions are increasing in s, with $\pi_T(s) > \pi_N(s)$ for all $s > 0$ (the system performs better than random). The cost of surveillance is $C_G(s, t)$, decreasing in technology level t.

We assume $C_G(s, t)$ is increasing and strictly convex in s, and that higher technology lowers the marginal cost of surveillance:

$$\frac{\partial C_G}{\partial s}(s, t) > 0, \qquad \frac{\partial^2 C_G}{\partial s^2}(s, t) > 0, \qquad \frac{\partial^2 C_G}{\partial s\, \partial t}(s, t) \leq 0.$$

Population Response. Individuals choose adaptation effort $a_\theta \in [0, 1]$. As seen in the literature, adaptation includes avoiding monitored areas, limiting social connections, modifying communication patterns, and other behavioral changes that reduce detectability. For type T individuals, adaptation reduces the probability of detection. For type N individuals, adaptation may reduce false positive risk but imposes costs on daily life, such as the shrinking of their private spaces.

Let $C_P(a)$ denote the cost of adaptation, increasing and strictly convex:

$$C_P'(a) > 0, \qquad C_P''(a) > 0.$$

Probability bounds. We interpret $\pi_T(\cdot)$ and $\pi_N(\cdot)$ as probabilities. To ensure they are always well-defined and lie in $[0, 1]$, we restrict choices so that the argument of these functions never leaves their domain. Let $\pi_T, \pi_N : [0, \bar{s}] \to [0, 1]$ be increasing and concave. We also bound adaptation to $a_\theta \in [0, 1]$ for each type $\theta \in \{T, N\}$. Then for any feasible (s, a_θ) with $s \in [0, \bar{s}]$,

$$s(1 - a_\theta) \in [0, \bar{s}],$$

so $\pi_T\big(s(1 - a_T)\big)$ and $\pi_N\big(s(1 - a_N)\big)$ are always valid probabilities. In this way, adaptation is modeled as a fractional reduction in "effective surveillance," so it can scale detectability down but cannot make it negative or push the model outside the range where the probability functions are defined.

4.2 Payoffs

Government Payoffs. The government receives benefit B from correctly identifying threats (preventing attacks, demonstrating control) and incurs cost F from false positives (reputational damage, legal challenges, resistance to occupation). However, following the empirical observation that false positives may serve political functions in population control, we introduce parameter $\alpha \in [0, 1]$ representing the political instrumentalization of errors. When $\alpha > 0$, the government derives partial benefit from false positives because they contribute to the general climate of fear and control.

Government expected payoff:

$$U_G = \lambda \cdot B \cdot \pi_T\big(s(1 - a_T)\big) + (1 - \lambda) \cdot [\alpha B - (1 - \alpha)F] \cdot \pi_N\big(s(1 - a_N)\big) - C_G(s, t) \tag{1}$$

where a_T and a_N denote adaptation levels of threat and non-threat types respectively.

Population Payoffs. Individuals face costs from being flagged (detention, harassment, restriction of movement) and from adaptation efforts. Type T individuals additionally value successful evasion of detection. For simplicity, assume:

$$U_T = V_T\big(1 - \pi_T\big(s(1 - a_T)\big)\big) - D \cdot \pi_T\big(s(1 - a_T)\big) - C_P(a_T) \tag{2}$$

$$U_N = -D \cdot \pi_N\big(s(1 - a_N)\big) - C_P(a_N) \tag{3}$$

where V_T is the value to type T of evading detection and D is the cost of being flagged by the system.

4.3 Timing and Equilibrium Concept

The game has two stages. In Stage 0, a surveillance technology level t is deployed (taken as exogenous) and is publicly observed; it affects the cost of surveillance through $C_G(s, t)$. In Stage 1, given t, the government chooses surveillance intensity s and individuals choose adaptation effort a simultaneously; each individual observes their own type $\theta \in \{T, N\}$, but types are not publicly observed.

Definition 1 (Bayesian Nash equilibrium conditional on technology)**.** Fix t. A Bayesian Nash equilibrium of the Stage 1 game induced by t is a triple $(s^*(t), a_T^*(t), a_N^*(t))$ such that:

(i) $s^*(t)$ maximizes the government's expected payoff given $(a_T^*(t), a_N^*(t))$;

(ii) for each type $\theta \in \{T, N\}$, $a_\theta^*(t)$ maximizes that type's expected payoff given $s^*(t)$.

In what follows, we characterize the Stage 1 Bayesian Nash equilibrium for a given t and derive comparative statics with respect to t (and α).

5 Analysis

The first-order conditions for optimal choices yield reaction functions. For the government, we set the function as follows:

$$\frac{\partial U_G}{\partial s} = \lambda B\, \pi_T'\big(s(1-a_T)\big)(1-a_T) + (1-\lambda)[\alpha B - (1-\alpha)F]\, \pi_N'\big(s(1-a_N)\big)(1-a_N) - \frac{\partial C_G}{\partial s} = 0 \tag{4}$$

For type T individuals:

$$\frac{\partial U_T}{\partial a_T} = (V_T + D)\, s\, \pi_T'\big(s(1 - a_T)\big) - C_P'(a_T) = 0 \tag{5}$$

For type N individuals:

$$\frac{\partial U_N}{\partial a_N} = D\, s\, \pi_N'\big(s(1 - a_N)\big) - C_P'(a_N) = 0 \tag{6}$$

Proposition 1 (Existence (and uniqueness under strict concavity)). *Fix t. If strategy sets are compact and payoffs are continuous in (s, a_T, a_N), a Bayesian Nash equilibrium exists. If, in addition, each player's payoff is strictly concave in its own action (for instance, under the curvature-dominance conditions stated in the Appendix), then the equilibrium is unique. In common parametric specifications, the unique equilibrium is interior.*

Existence follows from continuity of payoffs and compactness of the strategy spaces, which ensure a fixed point of best responses. The strategic interaction between surveillance and adaptation is driven by the fact that higher surveillance increases the stakes of being detected or falsely flagged, raising incentives to adapt, while adaptation reduces effective detectability through the argument $s(1 - a_\theta)$ in $\pi_\theta(\cdot)$, shaping the government's optimal surveillance choice.

Proposition 2 (Technology Increases Surveillance). *In equilibrium, surveillance intensity s^* is increasing in technology level t (holding the population cost function C_P fixed).*

This extends Dragu and Lupu (2021)'s core finding to the incomplete information setting. As technology reduces the cost of surveillance, the government optimally expands monitoring even when the population adapts in response. The strategic interaction between surveillance and adaptation creates an 'arms race' dynamic, but the government consistently gains in equilibrium from a commitment/first-mover advantage in deploying t (Stage 0), which lowers marginal surveillance cost in Stage 1.

Proposition 3 (Error Instrumentalization Expands Surveillance). *Equilibrium surveillance s^* is increasing in the instrumentalization parameter α. As false positives become more politically valuable, the government expands surveillance beyond what would be optimal for pure security purposes. Individuals' payoffs do not depend on α, so the only way α moves equilibrium is through the government's best response.*

Increasing α raises the marginal payoff weight on false positives in the government's objective, strengthening the incentive to increase s because $\pi_N\big(s(1 - a_N)\big)$ rises with surveillance. When $\alpha = 0$, the government faces a standard tradeoff between security benefits and false positive costs. But when $\alpha > 0$, false positives partially benefit the government by contributing to population control. In the limit as $\alpha \to 1$, every flagging, whether accurate or not, serves the government's interests. The comparative static result on α formalizes the observation from Hebron that algorithmic errors are not simply tolerated but may be functionally integrated into the surveillance system.

Proposition 4 (Population Welfare Declines with Technology). *Expected utility of both type T and type N individuals is decreasing in technology level t and in instrumentalization parameter α.*

Equilibrium utility decreases because equilibrium surveillance increases, and utility is decreasing in surveillance even after best-responding in adaptation.

A higher t induces higher equilibrium surveillance, increasing both true-positive detection of type T and false-positive flagging of type N, while also increasing incentives to invest in costly adaptation. A higher α further strengthens the government's incentive to expand surveillance even when it produces false positives, worsening expected outcomes for non-threats in particular. Under the model's assumptions, equilibrium welfare for both types weakly decreases with surveillance capability, and may decline strictly in typical cases: type T individuals face higher detection probability, while type N individuals face more frequent false positives and must expend more resources on adaptation. The burden falls disproportionately on non-threats, who gain nothing from successful evasion but bear costs of both flagging and adaptation. Moreover, when $\alpha > 0$, the welfare calculus becomes zero-sum between government and population. The government's political benefit from false positives is precisely the population's suffering from unjust flagging. This formalizes the sense in which algorithmic surveillance in contexts like Hebron constitutes what Goodfriend (2023) terms 'algorithmic state violence.'

6 Discussion, Policy Implications, and Limitations

The strategic interaction between surveillance and adaptation ensures that technological improvement does not reduce the intrusiveness of control but rather shifts its form. In equilibrium, as surveillance technology improves, the population adapts more intensively. This adaptation, including avoiding certain areas, limiting social connections, and modifying daily routines, represents a diffuse form of control that operates through self-discipline rather than direct coercion. Standard discussions of algorithmic governance treat errors as problems to be minimized through better data and improved algorithms. The analysis here suggests a different interpretation: when surveillance serves political functions beyond security, errors may become features that enhance rather than undermine the system's effectiveness. In our model, the government's net payoff weight on false positives is

$$k(\alpha) \equiv \alpha B - (1-\alpha)F = \alpha(B+F) - F.$$

This implies a critical threshold

$$\alpha^* = \frac{F}{B+F}$$

such that when $\alpha < \alpha^*$ false positives are net costly at the margin (the standard "accuracy" logic), while when $\alpha > \alpha^*$ false positives become net beneficial at the margin (an "errors as control" regime). This provides a clean political-economy restatement of the Hebron case: where reputational or legal constraints are weak (low F), the threshold α^* falls, making it easier for the system to rationally tolerate or even prefer "messy" outcomes. In such settings, the uncertainty produced by misclassification becomes a mechanism of generalized discipline: when civilians cannot predict whether they will be correctly or incorrectly classified, they must adapt their behavior as if they might always be flagged (Goodfriend, 2023). A related implication is that expanded surveillance can persist even when the prevalence of true threats is low. Even if λ

is small, a sufficiently large α (so that $k(\alpha) > 0$) gives the government a continuing incentive to maintain or expand surveillance because control benefits arise from broad flagging of the population, not only from true positives. This formalizes the intuition that in contexts of occupation and population management, the logic of control can dominate the logic of security.

The model also clarifies which regulatory levers matter. International scrutiny, domestic judicial review, media attention, and other forms of oversight effectively raise the expected cost of false positives F (or, equivalently, reduce the feasibility of instrumentalization by lowering the effective α). The comparative prediction is straightforward: stronger accountability should reduce equilibrium surveillance intensity and reduce the tolerance for false positives, while weaker accountability permits higher s^* and greater reliance on false positives as a tool of control. This connects naturally back to Dragu and Lupu (2021)'s emphasis that institutional and legal constraints can be conceptualized as increasing the costs of preventive control. In this sense, the "accuracy" of the algorithm is not the central policy variable; the binding constraint is whether political and legal institutions make errors costly.

A further implication is that the effective degree of instrumentalization need not be set only at the top of the hierarchy. The empirical record describes soldier photo quotas and competitive incentives for achieving 'pairings' (Kubovich, 2022). This suggests a principal–agent extension: implementers may receive local benefits from flagging and data production (recognition, rewards, career incentives), effectively increasing the perceived benefit term associated with flagging (raising an implementer-level analogue of α or B) even when central authorities bear some reputational cost F. As a result, even if leadership prefers higher accuracy, street-level incentives can push the system toward a regime in which false positives are operationally convenient and politically tolerable. In this interpretation, "errors as features" can emerge endogenously from incentive design and institutional structure, not only from explicit political intent.

Our model also has some evident limitations. The first is that it is static, while the Blue Wolf context and the broader literature emphasize dynamic feedback loops and learning over time. A minimal dynamic extension would allow technology t (or the performance parameters embedded in π_T and π_N) to improve with data volume, and data volume to rise with surveillance intensity s. This might yield a self-reinforcing escalation of surveillance: higher s generates more data (through increased encounters, images, and registrations), which improves operational capacity and lowers marginal costs, which then sustains or increases s in the next period. This mechanism is consistent with the reported emphasis on data accumulation and the gamification of collection in Hebron (Kubovich, 2022; Goodfriend, 2023). Moreover, the model treats the population as atomistic individuals, abstracting from collective action and organized resistance. Extending the framework to incorporate coordination among surveilled populations would illuminate additional dynamics.

7 Possible Expansion: Signaling, Thresholds, and Institutional Accountability

A natural extension is to recast the Hebron setting as a signaling environment in which Palestinians strategically choose observable behaviors that act as signals, while the state combines

these signals with algorithmic outputs to decide whether to detain or allow passage. This connects the paper's incomplete-information logic to classic signaling models (e.g., Beer–Quiche), but modifies them in three empirically relevant ways: (i) both types may prefer to appear "non-threatening"; (ii) the receiver's decision rule is mediated by a noisy algorithmic score with an adjustable classification threshold; and (iii) political accountability and street-level incentives shape how costly (or beneficial) errors are for the state.

7.1 From Beer–Quiche to "Protest–Silence"

Consider a representative civilian (sender) with private type $\theta \in \{T, N\}$, where T denotes "threat" and N denotes "non-threat." The sender chooses an observable signal $x \in \{L, H\}$. Interpret H as a politically salient or risk-associated behavioral profile (e.g., participation in protests, dense social ties, frequent presence in monitored areas), and L as a low-visibility profile (avoidance, self-silencing, cautious movement).

Unlike standard Beer–Quiche, it is plausible here that both types strictly prefer to be treated as "civilian" because being classified as a threat is costly regardless of θ. Thus, separation need not arise from a type-specific "taste" for a risky signal; instead, the equilibrium may feature pooling on L (self-silencing) because H becomes dangerous for everyone once it is surveilled and encoded as a risk feature.

7.2 Technology affects both cost and accuracy

In the baseline model, technology t lowers the marginal cost of surveillance through $C_G(s, t)$. In the signaling extension, it is also natural to let technology shift the informativeness of the score y.

A convenient "accuracy channel" assumption is that higher t improves type separation in the score distributions. One reduced-form way to state this is:

$$\frac{\partial \pi_T}{\partial t}(s, \tau, t \mid x) \geq 0, \qquad \frac{\partial \pi_N}{\partial t}(s, \tau, t \mid x) \leq 0,$$

for relevant (s, τ, x) (i.e., better technology raises detection of threats while lowering false positives at a given threshold), though weaker assumptions can be used (e.g., a monotone-likelihood-ratio property in (y, θ) that becomes stronger with t).

This two-channel formulation is substantively useful because it breaks the mechanical link "better technology $\Rightarrow$ more surveillance." When t increases accuracy, the state could in principle reduce s or raise τ (less harsh policy) while preserving detection performance, but whether that happens depends on the state's incentives over false positives.

7.3 Algorithm as a noisy decoder and the government's threshold choice

Let the state (receiver) observe the chosen signal x and an algorithmic risk score $y \in \mathbb{R}$ produced by the surveillance system. The score distribution depends on type, signal, surveillance intensity, and technology:

$$y \sim f_\theta(\cdot \mid x, s, t).$$

The receiver then chooses an action $r \in \{0, 1\}$, where $r = 1$ denotes "flag/detain" and $r = 0$ denotes "clear/release."

A key addition is that the state chooses a classification threshold $\tau \in \mathbb{R}$ (or, equivalently, a mapping from colors to actions). A simple decision rule is:

$$r(y; \tau) = \mathbf{1}\{y \geq \tau\}.$$

This explicitly separates (i) deployment intensity s (how much surveillance is applied) from (ii) policy strictness τ (how aggressive the labeling rule is). For each signal x, define the induced true-positive and false-positive rates:

$$\pi_T(s, \tau, t \mid x) = \Pr(y \geq \tau \mid \theta = T, x, s, t), \qquad \pi_N(s, \tau, t \mid x) = \Pr(y \geq \tau \mid \theta = N, x, s, t).$$

Holding (s, t, x) fixed, lowering τ increases both π_T and π_N. In words: a more aggressive threshold detains more true threats but also increases false positives among civilians.

This threshold choice makes the paper's substantive "harsher policy" claim literal: harsher governance can manifest either as larger s (more encounters, more scans) or as lower τ (a more aggressive rule that converts scores into detentions).

7.4 Accountability, threat prevalence, and a persistence condition

Let $\lambda \in (0, 1)$ denote the prior probability of type T. As in the main model, let B be the benefit from correctly flagging threats and F the cost of false positives. Let $\alpha \in [0, 1]$ capture the extent to which false positives are politically instrumentalized. Define the net marginal weight on false positives:

$$k(\alpha) = \alpha B - (1 - \alpha)F = \alpha(B + F) - F.$$

In the extended model, the government's expected objective can be written schematically as a function of (s, τ):

$$U_G(s, \tau; t) = \lambda B\, \pi_T(s, \tau, t \mid x) + (1 - \lambda)k(\alpha)\, \pi_N(s, \tau, t \mid x) - C_G(s, t),$$

where the dependence on x reflects the sender's equilibrium signaling behavior (pooling or separating).

This representation highlights three content-relevant comparative statics:

(i) Accountability as F. Holding α fixed, increasing F lowers $k(\alpha)$ and reduces the marginal attractiveness of policies that generate false positives. Thus, stronger oversight (higher expected F) should push toward lower surveillance intensity s and/or a less aggressive threshold (higher τ), even if technology is advanced.

(ii) Threat prevalence as λ. Increasing λ raises the marginal value of true positives and tends to increase s and/or lower τ. However, when α is large, policy becomes less sensitive to λ because a larger share of the government's payoff comes from broad flagging rather than accurate targeting.

(iii) Surveillance persistence when threats are rare. A sharp implication emerges from the sign of $k(\alpha)$. If $k(\alpha) > 0$ (equivalently, $\alpha > \alpha^* = \frac{F}{B+F}$), then false positives are

beneficial at the margin. In that regime, the government can have an incentive to maintain a harsh policy even when λ is very small: as $\lambda \to 0$, the false-positive term remains, so optimal (s, τ) need not converge to minimal surveillance. Substantively, this formalizes the idea that control incentives can sustain surveillance even when genuine threats are rare.

8 Conclusion

This paper has developed a game-theoretic framework to analyze algorithmic surveillance as a strategic interaction between a state authority and a targeted population under incomplete information. Using Israel's Blue Wolf system in Hebron as an empirical case, we extended the preventive repression model of Dragu and Lupu (2021) to incorporate classification errors and behavioral adaptation by surveilled individuals. Three main findings emerge from the analysis. First, technological improvement in surveillance capacity unambiguously increases equilibrium monitoring intensity, even when the population responds with costly adaptation. Second, when false positives carry political value for population control, captured by the instrumentalization parameter α, the government rationally expands surveillance beyond what pure security objectives would warrant. Third, welfare losses from surveillance fall on both threat and non-threat types, but disproportionately burden ordinary civilians who bear the costs of flagging and adaptation without any offsetting benefit from evasion. The critical threshold $\alpha^* = F/(B + F)$ provides a simple diagnostic: where reputational or legal costs of errors are low, algorithmic systems can drift into an "errors as features" regime in which misclassification becomes functional rather than accidental. This reframes the policy problem. The challenge is not primarily technical, improving algorithmic accuracy, but institutional, ensuring that political and legal structures make errors costly. In contexts of occupation and weak accountability, the model predicts that surveillance will expand and persist even when the prevalence of genuine threats is low, because the logic of control dominates the logic of security. Several avenues for further research remain open. A dynamic extension could formalize how data accumulation and algorithmic learning create path-dependent escalation. Incorporating collective action among the surveilled population would illuminate the conditions under which coordinated resistance can shift equilibrium outcomes.

References

Amnesty International. (2023). *Automated apartheid: How facial recognition fragments, segregates and controls Palestinians in the OPT.* Amnesty International. Retrieved from https://www.amnesty.org/en/documents/mde15/6701/2023/en/

Abraham, Y. (2024, April 3). 'Lavender': The AI machine directing Israel's bombing spree in Gaza. *+972 Magazine.* Retrieved from https://www.972mag.com/lavender-ai-israeli-army-gaza/

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim Code.* Polity Press.

Benjamin, G. (2019). Playing at control: Writing surveillance in/for gamified society. *Surveillance & Society, 17*(5), 699–713.

Browne, S. (2015). *Dark matters: On the surveillance of Blackness.* Duke University Press.

Byler, D. (2022). *Terror capitalism: Uyghur dispossession and masculinity in a Chinese city.* Duke University Press.

Dayan, H. (2022). Israel/Palestine: Authoritarian practices in the context of a dual state crisis. In O. Topak, M. Mekouar, & F. Cavatorta (Eds.), *New authoritarian practices in the Middle East and North Africa* (pp. 131–151). Edinburgh University Press.

Dragu, T., & Lupu, Y. (2021). Digital authoritarianism and the future of human rights. *International Organization, 75*(4), 991–1017.

Dwoskin, E. (2021, November 8). Israel escalates surveillance of Palestinians with facial recognition program in West Bank. *The Washington Post.* Retrieved from https://www.washingtonpost.com/world/middle_east/israel-palestinians-surveillance-facial-recognition/2021/11/05/3787bf42-26b2-11ec-8739-5cb6aba30a30_story.html

Fatafta, M., & Nashif, N. (2017, October 23). Surveillance of Palestinians and the fight for digital rights (Policy brief). *Al-Shabaka: The Palestinian Policy Network.* Retrieved from https://al-shabaka.org/briefs/surveillance-of-palestinians-and-the-fight-for-digital-rights/

Gidron, N. (2023). Why Israeli democracy is in crisis. *Journal of Democracy, 34*(3), 33–45.

Goodfriend, S. (2023). Algorithmic state violence: Automated surveillance and Palestinian dispossession in Hebron's Old City. *International Journal of Middle East Studies, 55*, 461–478.

Harcourt, B. (2007). *Against prediction: Profiling, policing, and punishing in an actuarial age.* University of Chicago Press.

Jamal, A. (2007). Nationalizing states and the constitution of "hollow citizenship": Israel and its Palestinian citizens. *Ethnopolitics, 6*(4), 471–493.

Kubovich, Y. (2022, March 24). Israeli troops' new quota: Add 50 Palestinians to tracking database every shift. *Haaretz*. Retrieved from https://www.haaretz.com/israel-news/2022-03-24/ty-article/.premium/soldiers-not-allowed-off-shifts-until-they-enter-50-palestinian-names-in-database/00000180-5ba7-d97e-a7fb-7bf7361c0000

Shalhoub-Kevorkian, N. (2015). *Security theology, surveillance and the politics of fear*. Cambridge University Press.

Smooha, S. (2002). The model of ethnic democracy: Israel as a Jewish and democratic state. *Nations and Nationalism, 8*(4), 475–503.

Rouhana, N. N. (1997). *Palestinian citizens in an ethnic Jewish state: Identities in conflict*. Yale University Press.

Peled, Y. (2008). The evolution of Israeli citizenship: An overview. *Citizenship Studies, 12*(3), 335–345.

Zureik, E. (2011). Colonialism, surveillance, and population control: Israel/Palestine. In E. Zureik, D. Lyon, & Y. Abu-Laban (Eds.), *Surveillance and control in Israel/Palestine*. Routledge.

A Appendix: Notation and Proof Sketches

This appendix provides a summary of notation and a set of proof sketches for the formal results stated in Section 5. The goal is to record the key steps and the main sufficient conditions under which the claims hold, without reproducing full-length derivations. A complete set of proofs (with all regularity assumptions stated explicitly, boundary cases treated, and all steps written out) can be provided upon request.

A.1 Notation

Symbol	Meaning
$s \in [0, \bar{s}]$	surveillance intensity chosen by the government
t	technology level (higher t lowers marginal surveillance cost)
$\theta \in \{T, N\}$	individual type: threat (T) or non-threat (N)
λ	population share of type T
$a_T, a_N \in [0, 1]$	adaptation effort by types T and N
$\pi_T(\cdot)$	true-positive probability (detecting T correctly)
$\pi_N(\cdot)$	false-positive probability (flagging N incorrectly)
B	benefit from correctly detecting threats
F	cost of false positives (reputation, legal constraints, backlash)
$\alpha \in [0, 1]$	degree to which false positives are politically instrumentalized
$k(\alpha)$	net marginal weight on false positives: $k(\alpha) = \alpha B - (1 - \alpha)F$
D	individual cost of being flagged
V_T	type-T value of evasion (avoiding detection)
$C_G(s, t)$	government surveillance cost
$C_P(a)$	adaptation cost

A.2 Standing assumptions used in the sketches

The sketches below use the following standard assumptions, which are consistent with the model statements in Sections 4–5.

Assumption 1 (Feasible sets and smoothness)**.** The strategy sets are compact:

$$s \in [0, \bar{s}], \qquad a_T \in [0, 1], \qquad a_N \in [0, 1].$$

The functions $C_G(\cdot, \cdot)$ and $C_P(\cdot)$ are twice continuously differentiable, and $\pi_T(\cdot)$ and $\pi_N(\cdot)$ are twice continuously differentiable on $[0, \bar{s}]$.

Assumption 2 (Monotonicity and curvature)**.** The probability functions are increasing and concave:

$$\pi_T'(x) \geq 0, \quad \pi_N'(x) \geq 0, \quad \pi_T''(x) \leq 0, \quad \pi_N''(x) \leq 0, \qquad \text{for all } x \in [0, \bar{s}].$$

The cost functions are increasing and strictly convex in the relevant choices:

$$\frac{\partial C_G}{\partial s}(s,t) > 0, \qquad \frac{\partial^2 C_G}{\partial s^2}(s,t) > 0, \qquad C_P'(a) > 0, \qquad C_P''(a) > 0.$$

Assumption 3 (Technology reduces marginal surveillance cost)**.** Technology lowers the marginal cost of surveillance in the sense of a weakly negative cross-partial:

$$\frac{\partial^2 C_G}{\partial s\,\partial t}(s,t) \le 0.$$

A.3 Existence (and typical uniqueness)

Proof sketch (Existence and uniqueness under strict concavity). Fix t and consider the Stage 1 game. By compactness of the strategy sets and continuity of payoffs in (s, a_T, a_N), existence of a mixed-strategy equilibrium follows from standard fixed-point arguments. To obtain a pure-strategy equilibrium with well-behaved comparative statics, it is convenient to impose sufficient conditions guaranteeing single-valued best responses.

A sufficient route is strict concavity of each player's payoff in its own action. For the government, holding (a_T, a_N) fixed, the objective is

$$U_G(s; a_T, a_N, t) = \lambda B\,\pi_T\big(s(1-a_T)\big) + (1-\lambda)k(\alpha)\,\pi_N\big(s(1-a_N)\big) - C_G(s,t),$$

where

$$k(\alpha) = \alpha(B+F) - F.$$

Concavity of π_T and π_N implies the benefit terms are concave in s, while strict convexity of C_G implies $-C_G$ is strictly concave. A sufficient "curvature dominance" condition ensuring strict concavity of U_G in s is that for all feasible (s, a_T, a_N),

$$\frac{\partial^2 C_G}{\partial s^2}(s,t) > \lambda B(1-a_T)^2\big|\pi_T''(s(1-a_T))\big| + (1-\lambda)\big|k(\alpha)\big|(1-a_N)^2\big|\pi_N''(s(1-a_N))\big|.$$

Similarly, for type T the payoff is

$$U_T(s, a_T) = V_T - (V_T + D)\,\pi_T\big(s(1-a_T)\big) - C_P(a_T),$$

and strict concavity in a_T is ensured if

$$C_P''(a_T) > (V_T + D)s^2\big|\pi_T''(s(1-a_T))\big|.$$

For type N,

$$U_N(s, a_N) = -D\,\pi_N\big(s(1-a_N)\big) - C_P(a_N),$$

and strict concavity in a_N is ensured if

$$C_P''(a_N) > Ds^2\big|\pi_N''(s(1-a_N))\big|.$$

Under these conditions, each best response is single-valued and continuous. The joint best-

response map from the compact convex set $[0, \bar{s}] \times [0,1] \times [0,1]$ to itself has a fixed point, yielding a pure-strategy Bayesian Nash equilibrium. When best responses are single-valued, the equilibrium is unique. Interiorness follows when the unique optimizer satisfies the first-order conditions with strict inequalities ruling out boundary solutions (e.g., Inada-type behavior or sufficiently steep marginal costs near the boundaries). □

A.4 A useful monotonicity lemma

Lemma 1 (Monotonicity preserved under maximization)**.** *Fix a type $\theta \in \{T, N\}$. Suppose that for all feasible $a_\theta \in [0,1]$, the payoff $U_\theta(s, a_\theta)$ is weakly decreasing in s. Define the value function*

$$\widehat{U}_\theta(s) = \max_{a_\theta \in [0,1]} U_\theta(s, a_\theta).$$

Then $\widehat{U}_\theta(s)$ is weakly decreasing in s.

Proof. Let $s_2 > s_1$. For any feasible a_θ,

$$U_\theta(s_2, a_\theta) \le U_\theta(s_1, a_\theta).$$

Taking maxima over $a_\theta \in [0,1]$ on both sides yields

$$\max_{a_\theta \in [0,1]} U_\theta(s_2, a_\theta) \le \max_{a_\theta \in [0,1]} U_\theta(s_1, a_\theta),$$

which proves the claim. □

A.5 Comparative statics in t and α

Proof sketch (Technology increases surveillance). Fix (a_T, a_N) and write the government's payoff as

$$U_G(s; a_T, a_N, t) = \lambda B\, \pi_T\big(s(1-a_T)\big) + (1-\lambda) k(\alpha)\, \pi_N\big(s(1-a_N)\big) - C_G(s,t).$$

The benefit terms do not depend on t. Under the technology condition

$$\frac{\partial^2 C_G}{\partial s\, \partial t}(s,t) \le 0,$$

the function $-C_G(s,t)$ has increasing differences in (s,t), hence so does $U_G(s; a_T, a_N, t)$. Therefore the government's best response in s is nondecreasing in t (Topkis' monotonicity theorem). With uniqueness (or by selecting the smallest or largest best response), equilibrium surveillance $s^*(t)$ is increasing in t. □

Proof sketch (Error instrumentalization expands surveillance). The dependence on α enters only through the net weight on false positives

$$k(\alpha) = \alpha(B+F) - F.$$

Holding (a_T, a_N, t) fixed, differentiate U_G with respect to α:

$$\frac{\partial U_G}{\partial \alpha} = (1-\lambda)(B+F)\,\pi_N\big(s(1-a_N)\big).$$

Since $\pi_N(\cdot)$ is increasing and the argument $s(1-a_N)$ is increasing in s, the expression above is increasing in s. Thus U_G has increasing differences in (s, α), implying the government's best-response surveillance level is nondecreasing in α. With uniqueness (or a monotone selection), equilibrium surveillance s^* is increasing in α. □

Proof sketch (Population welfare declines with t and α). Fix (s, a_N). For a non-threat type,

$$U_N(s, a_N) = -D\,\pi_N\big(s(1-a_N)\big) - C_P(a_N),$$

so

$$\frac{\partial U_N}{\partial s} = -D\,\pi_N'\big(s(1-a_N)\big)\,(1-a_N) < 0.$$

Fix (s, a_T). For a threat type,

$$U_T(s, a_T) = V_T - (V_T + D)\,\pi_T\big(s(1-a_T)\big) - C_P(a_T),$$

so

$$\frac{\partial U_T}{\partial s} = -(V_T + D)\,\pi_T'\big(s(1-a_T)\big)\,(1-a_T) < 0.$$

Thus, for each type, payoffs are strictly decreasing in s for every fixed adaptation choice. By the lemma above, the maximized payoffs after best-responding in adaptation,

$$\widehat{U}_T(s) = \max_{a_T \in [0,1]} U_T(s, a_T), \qquad \widehat{U}_N(s) = \max_{a_N \in [0,1]} U_N(s, a_N),$$

are also decreasing in s.

From the previous comparative statics sketches, equilibrium surveillance s^* increases in technology t and in instrumentalization α. Therefore equilibrium expected utility of both types decreases in t and in α through the induced increase in s^*. □