DOI: https://doi.org/10.71317/kjard.2.11.2026.971

Kashmir Journal of Academic Research and Development

Journal homepage: https://rjsaonline.org/index.php/KJARD

Deep Machine Learning-Driven Intelligence Architectures for Precision Surgical Strike Operations: A Multidimensional Analysis of Algorithmic Targeting, Predictive Threat Assessment, and Strategic Transformation in Contemporary Asymmetric Warfare

Hafiza Saadia Shams
National Institute of Intelligence & Security Studies, Quaid-i-Azam University, Islamabad 45320, Pakistan
Email: saadiashams25@gmail.com
Dr. Muhammad Usman Malik
National Institute of Intelligence & Security Studies, Quaid-i-Azam University, Islamabad 45320, Pakistan
Email: usmanmalik836@gmail.com
Dr. Zulfqar Khan
National Institute of Intelligence & Security Studies, Quaid-i-Azam University, Islamabad 45320, Pakistan
Email: dirniiss@qau.edu.pk

ARTICLE INFO

Received:
September 28, 2026
Revised:
October 13, 2026
Accepted:
October 25, 2026
Available Online:
November 04, 2026

Keywords:
Algorithmic targeting; Deep learning; Precision strike; Predictive threat assessment; Partial identification; Meta-analysis; Human–machine teaming; Asymmetric warfare.

Corresponding Author:
Hafiza Saadia Shams
Email:
saadiashams25@gmail.com

ABSTRACT

Purpose: This article asks what open-source evidence can and cannot tell us about the claim that deep machine learning (DML) has made surgical strike operations more precise. It treats an intelligence architecture as a socio-technical chain running from data, through models and human validation, to operational and strategic outcomes, and it examines the weakest link in that chain: the contested ground truth on which both algorithms and their evaluators depend. Design/methodology/approach: We assemble a sourced workbook covering US strikes in Pakistan (2004–2017; 409 strikes, 3,096 deaths), four cross-theatre campaigns, six competing Pakistan casualty estimates, Airwars evidential grades for Somalia, and reported throughput of algorithmic targeting systems. The estimation strategy combines negative binomial count models with exposure offsets and sandwich errors, a Quandt–Andrews sup-LR break test with parametric bootstrap, BIC-weighted break-date posteriors, Dirichlet-multinomial composition models, Manski partial identification with Imbens–Manski intervals, randomization inference, Paule–Mandel random-effects meta-analysis with Hartung–Knapp–Sidik–Jonkman correction, and 100,000-draw Monte Carlo propagation of source ranges. Deep learning enters through a Poisson-loss deep neural ensemble, a long short-term memory (LSTM) forecaster and a label-noise variance decomposition; Erlang-C queueing models measure the human-review capacity implied by reported targeting throughput. Findings: Civilian deaths per strike fell by about 39 per cent a year, but the decline is a smooth trend rather than a datable regime break (sup-LR = 5.22, bootstrap p = 0.281), and it is not attributable to the 2013 Presidential Policy Guidance. Civilian deaths rose far less than proportionally with strike tempo (elasticity ≈ 0.25). Sources disagree so sharply that I^2 = 98.4 per cent and the 95 per cent prediction interval for the civilian share spans 0.1 to 93 per cent; label uncertainty accounts for 92.9 per cent of the deep ensemble's predictive variance. Deep models did not outperform parsimonious count models out of sample. At the reported rate of 1,000 recommendations an hour, a 20-person targeting cell can give each recommendation at most 1.2 minutes of review. Originality/value: The article is, to our knowledge, the first to combine partial identification, meta-analysis of competing casualty sources and deep-learning uncertainty decomposition in the study of algorithmic targeting, and to quantify the throughput–oversight constraint with queueing theory. Practical implications: Claims of algorithmic precision should be audited against bounded, multi-source ground truth, and review time per nomination should be treated as a measurable condition of meaningful human control.

Introduction

For most of the history of war, the hard part of striking an individual was not the weapon. It was knowing, with enough confidence to act, who the person was, where they were and whether anyone else would be hurt. Intelligence and analysis were rarely good enough to carry that burden, and commanders compensated with mass. The twentieth century narrowed the gap through aerial reconnaissance, signals intelligence and, eventually, persistent surveillance from remotely piloted aircraft (Vlad & Hardy, 2025). In the early 1990s Andrew Marshall's Office of Net Assessment argued that information superiority would sit at the centre of operational art, and that precision would follow from it (Marshall, 1993; Krepinevich, 1994; Cohen, 1996).

A second claim now sits on top of the first. Many scholars and practitioners argue that artificial intelligence (AI), and deep learning in particular, will change the character of war in the way that gunpowder, the tank and the aircraft did (Horowitz, 2018; King, 2024). States talk as if they believe it. China's 2017 development plan described AI as a strategic technology that would lead the future (Kania, 2017), and the US Department of Defense's first dedicated budget line for AI and autonomy, requested for fiscal year 2026, came to $13.4 billion. Project Maven, launched in 2017 to apply computer vision to full-motion video, became a programme of record in March 2026, and its reported user base grew from about 5,000 in early 2024 to 20,000 by May 2025. In Gaza, decision-support systems reported as "The Gospel" and "Lavender" have been described as producing target recommendations at a scale and speed that analysts working alone could not match (Davies et al., 2023; Abraham, 2024).

The language that accompanies these systems is the language of precision. Precision, however, is an empirical claim, and it has a peculiar structure. To show that an algorithm has made strikes more precise, one needs to know how many of the people killed were combatants and how many were not. That knowledge is exactly what is most disputed in covert and remote campaigns. In Pakistan, published tallies of civilian deaths from the same programme differ by a factor of about nineteen between the lowest and highest estimates. The problem is not only one for outside evaluators. A deep network trained to nominate targets or to forecast threats learns from labels produced by the same intelligence process, and battle-damage assessments that count unidentified adult males as militants feed back into the data on which future models are trained (Suchman, 2020; Schwarz, 2018). Precision can therefore be partly manufactured by the categories used to measure it.

This article takes that problem seriously and makes it the object of analysis. We use "deep machine learning-driven intelligence architecture" to mean the socio-technical assembly of sensors, data pipelines, learning models, human analysts, legal review and command authority through which algorithmic outputs become operational effects. AI in this sense is narrow AI: machine-learning systems, especially deep neural networks, that detect objects in imagery, fuse sensor streams, score entities or forecast events (LeCun et al., 2015; Goodfellow et al., 2016). We do not study weapons or their design. We study what the open record says about the outcomes such architectures are claimed to improve, and how much confidence that record can bear.

Three gaps motivate the study. First, the literature on military AI is mostly conceptual, normative or built on single cases; systematic evidence on whether outcomes associated with algorithmic intelligence actually changed is scarce. Second, quantitative work on targeted strikes (Johnston & Sarbahi, 2016; Jaeger & Siddique, 2018) generally treats one casualty source as the truth, so the uncertainty in the outcome variable itself never enters the analysis. Third, the throughput claims made for algorithmic systems have not been set against the arithmetic of human review, even though "meaningful human control" is now the central organizing concept of international policy debate (Santoni de Sio & van den Hoven, 2018; U.S. Department of State, 2023).

We ask four research questions. RQ1: Did civilian harm per strike decline over the Pakistan campaign, and if so, did it decline gradually or at an identifiable regime change? RQ2: How did civilian harm scale with strike tempo? RQ3: How large is the disagreement among casualty sources, and how much of the uncertainty in any learned model of strike outcomes comes from that disagreement rather than from the model? RQ4: Is the throughput reported for algorithmic targeting compatible with substantive human validation of each recommendation?

The article makes three contributions. Conceptually, it adds a label layer to the usual input–model–decision account of algorithmic targeting and shows why the reliability of outcome data is a strategic, not merely technical, variable. Methodologically, it brings together count-data econometrics, structural-break testing, partial identification, random-effects meta-analysis, Monte Carlo propagation, deep neural ensembles and queueing theory in one replicable pipeline written in Python. Substantively, it shows that the open record supports a large and real decline in reported civilian harm per strike, cannot attribute that decline to a specific technological or governance change, and is too contested for any learned model to settle the question. It also shows that reported throughput leaves very little time for human judgment.

The paper proceeds as follows. Section 2 reviews the literature and Section 3 sets out the framework and hypotheses. Section 4 describes the data and the estimators. Section 5 reports results, Section 6 discusses them, Section 7 draws out policy implications, Section 8 sets out limitations and Section 9 concludes.

Literature Review

Artificial Intelligence and Military Transformation

Students of international security have taken a sustained interest in AI's political, ethical and military consequences (King, 2024). Buchanan and Imbrie (2022) call AI "the new fire"; Kissinger et al. (2021) argue that it alters the epistemic foundations of strategy; Payne (2018, 2021) contends that it changes the psychological basis of strategic decision. Much of the early debate concerned lethal autonomous weapons (Arkin, 2009; Sparrow, 2007; Altmann & Sauer, 2017; Garcia, 2018; Bode & Huelss, 2018), with drone swarms the paradigm case (Scharre, 2018; Hambling, 2015). The debate has since moved towards systems actually in use (Horowitz, 2019; Johnson, 2021; Scharre, 2023; Brose, 2020). King (2024) distinguishes AI for data analysis, for decision support and for autonomous execution, and observes that its dominant contemporary use is the first two, not the third (see also Faris, 2025).

Work on military innovation warns against reading capability straight off technology. Military power depends on how technology is combined with doctrine, organization and force employment (Biddle, 2004; Horowitz, 2010; Mahnken, 2008), and diffusion is limited by absorptive capacity (Gilli & Gilli, 2019). Friction survives in information-rich battlespaces (Watts, 2004). These arguments justify our attention to the whole architecture rather than to algorithms alone.

Artificial Intelligence and Military Intelligence

Targeting has been the main proving ground for military AI. Algorithms are used to collect, collate and analyze data so that commanders can see the battlespace and act on it (Berman et al., 2018; Ford & Hoskins, 2022; González, 2022). Defence strategy now names data and analytics, rather than robotics, as the critical enabler for the coming decade (U.S. Department of Defense, 2023b). Goldfarb and Lindsay (2022) supply the most useful economic framing: machine learning reduces the cost of prediction but not the cost of judgment, so it raises rather than lowers the value of human decision-makers, especially when data are noisy and adversaries deceptive. Goldman and Grossman (2026) describe the resulting contest as analytic competition: a race to field analytics and to degrade the adversary's.

Algorithmic Targeting and the Construction of Precision

Deploying algorithmic systems in targeting is a strategic choice, not an accident of technical progress (Amoore, 2009, 2020). Reported accounts of Lavender describe a model that scored residents of Gaza against features associated with known combatants, using surveillance, communications and social-media data, with a self-assessed accuracy of about 90 per cent and a reported human review of roughly twenty seconds per target (Abraham, 2024). The earlier lineage runs through the targeted-killing programmes in Pakistan, Yemen and Somalia, built on persistent intelligence, surveillance and reconnaissance (ISR) and pattern-of-life analysis, including "signature strikes" against people whose behaviour matched a profile (Gregory, 2011; Chamayou, 2015; Vlad & Hardy, 2025).

Critics argue that algorithmic accuracy is constructed by the categories that define a target, so that technical precision can coexist with legal and political indeterminacy (Suchman, 2020; Schwarz, 2018). Defenders reply that better discrimination reduces harm relative to cruder alternatives (Byman, 2013); sceptics stress strategic costs (Cronin, 2013). Both sides lean on casualty numbers, and both tend to choose the source that suits them. That is the gap our label-layer analysis addresses.

Predictive Threat Assessment

Predictive threat assessment applies statistical learning to the question of where and when violence will occur. Political science has moved from explanatory models (Fearon & Laitin, 2003) to explicit forecasting, and has found that out-of-sample performance can diverge sharply from in-sample significance (Ward et al., 2010). Early-warning systems such as ViEWS (Hegre et al., 2019), text-based predictors (Mueller & Rauh, 2018) and point-process models (Mohler et al., 2011) show real but bounded skill, and Cederman and Weidmann (2017) argue that expectations should be adjusted to irreducible uncertainty. Deep architectures, from convolutional networks (Krizhevsky et al., 2012) to recurrent networks (Hochreiter & Schmidhuber, 1997) and transformers (Vaswani et al., 2017), add capacity but also new failure modes: adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015) and opacity in high-stakes settings (Rudin, 2019). On small tabular data, moreover, deep networks routinely lose to simpler learners (Grinsztajn et al., 2022), and in forecasting competitions statistical methods have often beaten machine-learning ones (Makridakis et al., 2018).

A further strand, less visible in security studies, concerns label noise. Supervised models inherit the errors of their labels, and when those errors are systematic rather than random, more data and larger models do not remove them (Frénay & Verleysen, 2014; Northcutt et al., 2021). Bayesian deep learning separates the uncertainty that comes from the data from the uncertainty

that comes from the model (Kendall & Gal, 2017; Lakshminarayanan et al., 2017). We use that idea to ask how much of what a strike-outcome model "does not know" is really disagreement among the people who count the dead.

Human–Machine Teaming and Automation Bias

Human–machine teaming (HMT) is a feedback process in which human and machine change each other's behaviour; humans bring context and judgment, machines bring speed and consistency (Konaev & Chahal, 2021). Project Maven is a standard example (Vincent, 2022). Human-factors research identifies how teaming fails: operators misuse, disuse or abuse automation (Parasuraman & Riley, 1997); complacency rises with workload and system reliability (Parasuraman & Manzey, 2010); in time-critical targeting, operators accept machine recommendations even against contrary evidence (Cummings, 2004). Calibrated trust requires transparency (Lee & See, 2004). These findings underpin the concept of meaningful human control (Santoni de Sio & van den Hoven, 2018; Roff, 2014) and its appearance in policy (U.S. Department of Defense, 2023a; International Committee of the Red Cross, 2021; United Nations General Assembly, 2023). What is missing is a way to put a number on the time that control requires.

Asymmetric Warfare, Decapitation and Adaptation

The weaker side in asymmetric wars often prevails by adopting an opposite strategic approach (Arreguín-Toft, 2001) or by outlasting the stronger side's resolve (Mack, 1975). Mechanization can weaken counterinsurgents by distancing them from the population (Lyall & Wilson, 2009), and intelligence flows from the population are decisive (Kalyvas, 2006; Berman et al., 2011). Evidence on leadership targeting is mixed (Price, 2012; Johnston, 2012; Jordan, 2009, 2014). For Pakistan, Johnston and Sarbahi (2016) associate strikes with lower militant violence, while Jaeger and Siddique (2018) find short-lived effects. Militants adapt by dispersing and changing communication patterns (Haas & Fischer, 2017), a dynamic that erodes the value of technically superior ISR.

Research Gap

Three gaps follow. The first is measurement: quantitative studies treat one casualty series as truth, and none, to our knowledge, carries source disagreement through to inference. The second is attribution: improvements in reported precision are routinely credited to technology or to policy without testing whether the data distinguish a regime change from a trend. The third is capacity: claims about algorithmic throughput and claims about human control are made in separate literatures and are rarely confronted with each other. The present study is designed to address all three.

Theoretical Framework and Hypotheses

We model a DML-driven intelligence architecture as a production chain with five layers (Figure 1). The input layer comprises ISR, signals, open-source and pattern-of-life data. The model layer converts inputs into detections, entity scores and threat forecasts. The decision layer is where humans validate recommendations under rules of engagement and legal review. Outputs appear first as operational outcomes, strike tempo and civilian harm, and then as strategic outcomes such as adversary adaptation, legitimacy and deterrence. We add a fifth, label layer: the battle-damage assessments and casualty counts that serve both as the evaluator's evidence and as the training data for the next generation of models. Two feedbacks matter. Outcomes become labels, and the throughput of the model layer sets the time available at the decision layer.

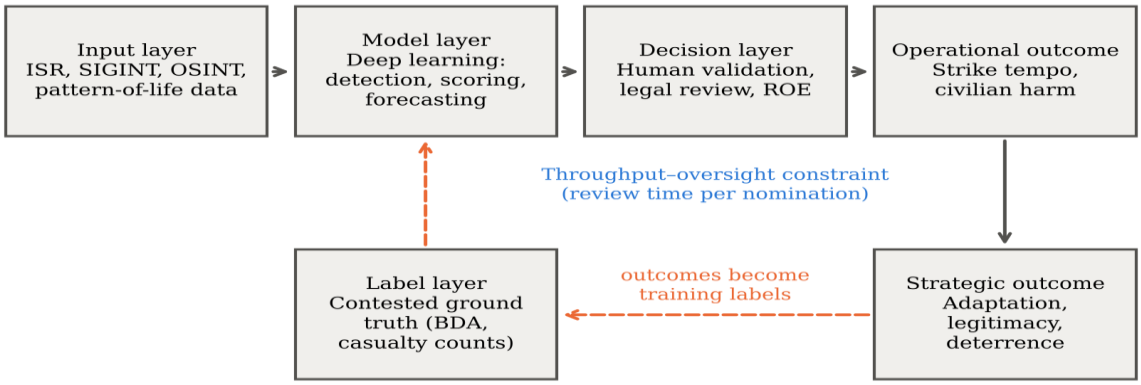


Figure 1. Conceptual framework: layers of a DML-driven intelligence architecture, the label feedback loop and the throughput–oversight constraint. Source: Authors' construction.

Learning Logic. If better sensors, fusion and models improved discrimination over the life of a campaign, civilian harm per strike should fall. Whether it falls gradually, as organizational learning and capability accumulate, or abruptly, at a governance or technology shock such as the May 2013 Presidential Policy Guidance (PPG) on direct action, is an empirical question with different implications for attribution.

H1: Civilian harm per strike declined over the Pakistan campaign, and the decline is better described by a gradual trend than by a discrete regime break.

Scaling Logic. If strikes were produced by a constant process, civilian deaths would scale one-for-one with the number of strikes. A less-than-proportional relationship would be consistent with learning by doing, with selection of easier targets in high-tempo years, or with reporting that thins out as strikes become routine. A more-than-proportional relationship would be consistent with the "industrialization" of targeting that critics fear.

H2: Civilian deaths do not scale proportionally with strike tempo.

Label Logic. If casualty sources disagree substantially and systematically, uncertainty about the outcome variable will exceed sampling and model uncertainty, and any learned model of strike outcomes will inherit that disagreement.

H3: Disagreement among casualty sources dominates model uncertainty in predicting civilian harm.

Capacity Logic. Deep networks lower the cost of prediction (Goldfarb & Lindsay, 2022) but expand the volume of predictions that humans must judge. The competing expectation from the deep-learning literature is that richer models extract more signal; the expectation from the small-data and forecasting literature is the reverse (Grinsztajn et al., 2022; Makridakis et al., 2018).

H4: In small, contested open-source data, deep neural architectures do not outperform parsimonious count-data models out of sample.

H5: The throughput reported for algorithmic targeting systems exceeds the capacity of reported human staffing to give each recommendation substantive review.

Data and Methodology

Data

The empirical base is a workbook compiled on 31 October 2026 from published open sources and documented sheet by sheet. Every hard-coded figure is traced to a cited source; derived quantities are calculated in the workbook and recalculated in our code. Table 1 lists the components. The core series is New America's (2018) annual record of US strikes in Pakistan, 2004–2017, covering 409 strikes and 3,096 deaths, of whom 2,533 were reported as militants, 288 as civilians and 275 as unknown. A further strike recorded in 2018 by the Bureau of Investigative Journalism (TBIJ) is excluded so that the series stays internally consistent. The cross-theatre data come from TBIJ's records (Bureau of Investigative Journalism, 2020) for Afghanistan, Pakistan, Somalia and Yemen (January 2004 to February 2020), reported as low–high ranges. The competing-estimates sheet records six assessments of civilian harm in Pakistan (New America, TBIJ, the Long War Journal, the Pakistan Ministry of Defence, a leaked Pakistani internal assessment and figures given by US officials), plus the 2016 statement of the Director of National Intelligence for all theatres outside areas of active hostilities (Office of the Director of National Intelligence, 2016). Airwars (2026) supplies civilian-harm allegations for Somalia by evidential grade. The remaining sheets record US defence AI budget lines (FY2024–FY2027), the institutional history of Project Maven, reported throughput of the Gospel and Lavender systems, SIPRI military expenditure for 2025 (Stockholm International Peace Research Institute, 2026) and vendor forecasts of the military AI market.

Table 1. Data Components, Coverage and Use

Component	Source(s)	Coverage	Used for
Pakistan annual strikes and deaths	New America	2004–2017; T = 14; 409 strikes	Count models, breaks, composition, bounds, deep learning
Cross-theatre campaigns	TBIJ	Afghanistan, Pakistan, Somalia, Yemen; 2004–Feb 2020	Monte Carlo propagation of ranges

Component	Source(s)	Coverage	Used for
Competing Pakistan estimates	New America, TBIJ, LWJ, Pakistan MoD, leaked Pakistani assessment, US officials, US DNI	Various, 2004–2020	Random-effects meta-analysis; label-noise design
Somalia evidential grades	Airwars	2007–2026; 558 incidents logged	Share of allegations meeting evidential thresholds
Algorithmic throughput	IDF statements, Guardian, NPR, +972/Local Call; NGA, DefenseScoop	2021–2026	Throughput ratios; Erlang-C staffing
Budgets and markets	DoD budget reporting; SIPRI; Visiongain; Persistence	FY2024–FY2027; 2025; 2026–2036	Context and diffusion indicators

Note: LWJ = Long War Journal; MoD = Ministry of Defence; DNI = Director of National Intelligence; NGA = National Geospatial-Intelligence Agency. Gaza and Maven throughput figures are reported by officials or journalists and have not been independently verified; they are treated as stated claims.

Two features of these data shape every choice that follows. The Pakistan series is short (fourteen annual observations), so large-sample panel methods are unavailable and every inference must respect small-sample constraints. And the outcome is contested at source, so the analysis must carry measurement disagreement forward rather than assume it away. We do not simulate any data; every estimate is computed from the published figures.

Variables

The main outcome is civilian deaths per year, C_t, modelled with strike count S_t as exposure. Secondary outcomes are the composition of deaths across militant, civilian and unknown categories, and the civilian share C_t/D_t, where D_t is total deaths. Regressors are a linear trend measured in decades from 2010, a PPG indicator equal to one from 2013, and centred log strike tempo. The trend proxies the cumulative maturation of ISR, data fusion and analytic capability over the campaign, which cannot be measured directly in open sources; the indicator captures the main governance change; tempo captures scale.

Econometric Strategy

We proceed from count models to tests of structural change, to models of composition and to bounds that make no assumption about the unknown category. Inference is then extended across sources and theatres, and finally to deep learning and to the capacity of human review.

Negative Binomial Count Models

Civilian deaths are non-negative integers with variance far above the mean (the variance-to-mean ratio is 47.1), so we use the NB2 model (Hausman et al., 1984; Cameron & Trivedi, 2013):

$$C_t \sim \mathrm{NB}(\mu_t, \alpha),\ \ \mu_t = \exp(\ln S_t + x_t'\beta),\ \ \mathrm{Var}(C_t) = \mu_t + \alpha\mu_t^2 \qquad (1)$$

The offset $\ln S_t$ makes $\exp(\beta)$ an incidence-rate ratio for civilian deaths per strike. Parameters are estimated by maximum likelihood. Because T is small and the model may be misspecified, we report HC1-scaled sandwich standard errors, $\hat{V} = H^{-1}(\Sigma\, s_t s_t')H^{-1} \cdot T/(T-k)$, where H is the Hessian and s_t the observation scores. Overdispersion is tested by a likelihood-ratio test of $\alpha = 0$ against the Poisson model, using the boundary-corrected $\tfrac{1}{2}\chi^2(1)$ reference distribution. When $\ln S_t$ also enters as a regressor with coefficient γ, the elasticity of civilian deaths with respect to tempo is $1 + \gamma$, so H2 is a test of $\gamma = 0$.

Structural break: Quandt–Andrews Sup-LR and Break-Date Posterior

To test H1 we estimate, for each candidate year τ, the model with an additional indicator $I(t \geq \tau)$ and compute $\mathrm{LR}(\tau)$ relative to the no-break trend model. Candidate dates run from 2007 to 2014, which trims roughly 15 per cent of the sample at each end. The test statistic is

$$\text{sup-LR} = \max_{\tau \in [2007, 2014]} 2[\ell_1(\tau) - \ell_0] \qquad (2)$$

(Quandt, 1960; Andrews, 1993; Bai & Perron, 1998). Because the break date is not identified under the null and T is small, asymptotic critical values are unreliable, so we obtain the null distribution by a parametric bootstrap: we simulate 299 series from the fitted no-break NB model (as gamma–Poisson mixtures) and recompute the full sup-LR each time. We also convert the Schwarz (1978) criterion of each model, including the no-break model, into approximate posterior probabilities, $w_m \propto \exp(-BIC_m/2)$, which gives a Bayesian model-averaged distribution over break dates.

Dirichlet-Multinomial Composition Model

The split of deaths among militants, civilians and unknowns is a vector of counts $n_t = (n_{1t}, n_{2t}, n_{3t})$. A multinomial logit would treat each death as an independent draw, which is implausible when deaths cluster by strike. We therefore estimate a Dirichlet-multinomial regression (Mosimann, 1962):

$$n_t \mid N_t \sim \text{DirMult}(N_t; a_{1t}, a_{2t}, a_{3t}),\ a_{jt} = \exp(x_t'\beta_j),\ E(n_{jt}/N_t) = a_{jt} / \Sigma_k a_{kt} \quad (3)$$

The precision $A = \Sigma_k a_k$ governs overdispersion; the intra-class correlation $1/(1 + A)$ measures how far deaths cluster beyond multinomial sampling. Average marginal effects on expected shares are computed numerically, with delta-method sandwich standard errors.

Partial Identification

The unknown category is not ignorable: it contains 8.9 per cent of all deaths. Rather than assigning it by assumption, we follow Manski (2003) and report the identified set for the civilian share,

$$\theta \in [\ C/D\ ,\ (C + U)/D\ ] \quad (4)$$

where U is unknown deaths. The lower bound assumes every unknown was a militant and the upper bound that every unknown was a civilian. For differences between periods, the identified set for $\theta_{post} - \theta_{pre}$ is $[L_{post} - U_{pre}, U_{post} - L_{pre}]$, and the sign of the change is identified only if this interval excludes zero. Confidence intervals for the identified set use the Imbens and Manski (2004) construction with year-level bootstrap standard errors (2,000 replications). For the period comparison in civilian deaths per strike we use randomization inference (Fisher, 1935): the observed difference in means is compared with its distribution under 20,000 random reassignments of years to periods, which is exact under the sharp null and requires no distributional assumption.

Random-Effects Meta-Analysis of Competing Sources

To test H3 at the source level we treat each independent assessment of the Pakistan civilian share as a study. Estimates are analysed on the logit scale. For sources reporting ranges, the logit of the midpoint share is the estimate and the reported range is treated as a 95 per cent interval, so $SE = [\text{logit}(\text{high}) - \text{logit}(\text{low})]/3.92$; for point estimates, $SE = (1/c + 1/(n - c))^{1/2}$. Between-source variance τ^2 is estimated by DerSimonian and Laird (1986) and by the iterative Paule and Mandel (1982) method, which performs better when the number of studies is small. With only four studies, normal-based intervals are far too narrow, so we apply the Hartung–Knapp–Sidik–Jonkman (HKSJ) correction with $t(k - 1)$ critical values (Hartung & Knapp, 2001):

$$\hat{\mu} = \Sigma w_i y_i / \Sigma w_i,\ w_i = 1/(v_i + \hat{\tau}^2),\ SE_HK = [\ \Sigma w_i (y_i - \hat{\mu})^2 / ((k - 1)\Sigma w_i)\ ]^{1/2} \quad (5)$$

Heterogeneity is summarized by Cochran's Q and by $I^2 = (Q - (k - 1))/Q$ (Higgins & Thompson, 2002), and the 95 per cent prediction interval, $\hat{\mu} \pm t_{k-2}(\hat{\tau}^2 + SE^2)^{1/2}$, shows where a new, equally credible source might fall. The DNI figure, which covers all theatres rather than Pakistan alone, is added only in a sensitivity analysis. Leave-one-out estimates measure the influence of each source.

Monte Carlo Propagation of Cross-Theatre Ranges

TBIJ reports civilian and total deaths as ranges. Treating midpoints as data hides that uncertainty. We draw civilian and total deaths independently and uniformly within each reported range, 100,000 times per theatre, enforce total ≥ civilian, and compute the distribution of the civilian share, deaths per strike and the child share of civilian deaths. From the joint draws we obtain the probability that each theatre has the highest civilian share and the distribution of the Spearman correlation between lethality per strike and civilian share across theatres.

Deep Learning

All neural models are implemented from first principles in NumPy so that every operation is transparent and reproducible. The first is a deep feed-forward network with three hidden layers (16, 16 and 8 units, tanh activations), dropout of 0.1 and L2 penalty

10^{-3}, trained with Adam (Kingma & Ba, 2015) for 3,000 epochs on a Poisson negative log-likelihood with the same exposure offset as equation (1):

$$L(\theta) = (1/T)\,\Sigma_t\,[\,\mu_t(\theta) - C_t \ln \mu_t(\theta)\,],\;\; \ln \mu_t(\theta) = \ln S_t + f_\theta(x_t) \quad (6)$$

Ten networks with different initializations form a deep ensemble (Lakshminarayanan et al., 2017). Predictive performance is assessed by leave-one-year-out cross-validation against the NB2 and Poisson versions of model M2 and an intercept-only NB benchmark, using Poisson deviance, mean absolute error and root mean squared log error. A Diebold and Mariano (1995) test compares the squared log errors of the ensemble and the NB2 model.

The second model is a long short-term memory network (Hochreiter & Schmidhuber, 1997) with eight hidden units, trained with back-propagation through time on rolling three-year windows of log strike counts. We produce one-step-ahead forecasts for 2011–2017 in a strict walk-forward design, retraining a five-member ensemble each year on data available up to that year, and compare it with a random walk and an AR(1) model estimated by least squares.

The third exercise is a label-noise experiment that speaks directly to H3. We generate 24 alternative label sets that are each consistent with the observed range of sources. Each set rescales the annual civilian series by a common factor drawn log-uniformly between the lowest and highest source totals (50 and 969 civilians, against 288 in the New America series) and multiplies each year by lognormal noise whose spread matches the width of TBIJ's reported range. On each label set we train five networks with different seeds. By the law of total variance,

$$\mathrm{Var}(\hat{y}) = E_L[\,\mathrm{Var}_\mathrm{seed}(\hat{y} \mid L)\,] + \mathrm{Var}_L[\,E_\mathrm{seed}(\hat{y} \mid L)\,] \quad (7)$$

the first term is model (epistemic) uncertainty and the second is label uncertainty. We also record, for every network, the counterfactual ratio of predicted civilian deaths with the PPG indicator switched on versus off.

Throughput and Human-Review Capacity

To test H5 we treat human validation of algorithmic recommendations as a queue. Recommendations arrive at rate λ, and each requires a mean review time s. With c analysts working in parallel, the offered load is $A = \lambda s$ and the probability that a recommendation must wait is given by the Erlang-C formula (Gross et al., 2008):

$$P_\mathrm{wait}(c, A) = c \cdot B(c, A) \,/\, [\,c - A(1 - B(c, A))\,] \quad (8)$$

where $B(c, A)$ is the Erlang-B blocking probability, obtained from the recursion $B(k, A) = A \cdot B(k-1, A)/[k + A \cdot B(k-1, A)]$ with $B(0, A) = 1$.

We compute the smallest c such that $P_\mathrm{wait} \le 0.2$ for $\lambda = 1{,}000$ recommendations per hour, the rate attributed to the Maven Smart System by the NGA Director, across review times from twenty seconds to twenty minutes. The point of the exercise is not operational. It is to translate a reported throughput into the human time it implies and to compare that with the reported size of a targeting cell. For the reported Lavender figures we compute the implied number of misidentified individuals at alternative accuracies and, treating the stated 90 per cent as the result of a validation sample of size n, the 95 per cent Beta-posterior interval for that number.

All code is in Python 3.13 (NumPy, SciPy, pandas, scikit-learn, Matplotlib) with fixed random seeds; scripts and the workbook are available as supplementary material.

Empirical Results

Descriptive Statistics

Table 2 summarizes the Pakistan series. Strikes rose from one in 2004 to a peak of 122 in 2010 and fell to three in 2016. Every series is strongly right-skewed and over dispersed: variance-to-mean ratios range from 38 for unknown deaths to 294 for total deaths, against the value of one implied by a Poisson process. The civilian series also carries one extreme observation. In 2006 two strikes are recorded as killing 93 civilians, most of them in a single strike on a religious school in Bajaur. That year alone contributes almost a third of all recorded civilian deaths, and we return to its influence repeatedly.

Table 2. Descriptive Statistics, US Strikes in Pakistan, 2004–2017

Variable	Mean	SD	Min	Max	Skewness	Kurtosis	Var/Mean
Strikes	29.21	34.77	1	122	1.48	4.55	41.38
Militants killed	180.93	224.55	1	788	1.53	4.74	278.70
Civilians killed	20.57	31.12	0	93	1.36	3.32	47.09
Unknown killed	19.64	27.44	0	92	1.43	4.29	38.34
Total killed	221.14	254.87	7	849	1.27	3.59	293.75

Note: T = 14 annual observations; totals 409 strikes, 2,533 militants, 288 civilians, 275 unknown, 3,096 deaths. Source: New America, compiled in the study workbook. Var/Mean = variance-to-mean ratio (1 under a Poisson process).

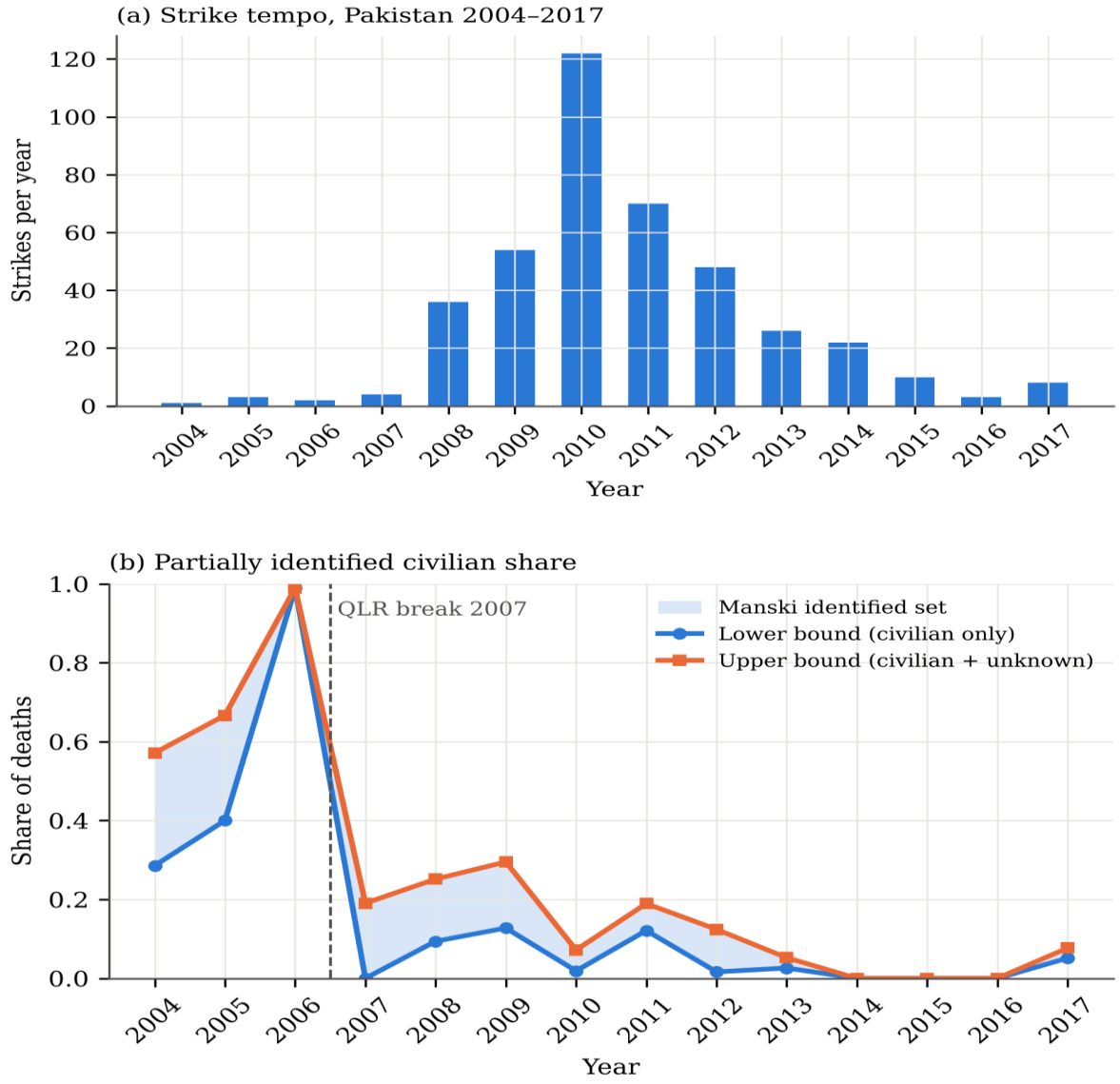


Figure 2. (a) Strike tempo and (b) the Manski identified set for the civilian share of deaths, Pakistan, 2004–2017. The lower bound counts only reported civilians; the upper bound also counts all deaths of unknown status. The dashed line marks the date that maximizes the sup-LR statistic.

Figure 2 shows the two features that every model must reconcile. Tempo is concentrated in 2008–2012, and the civilian share collapses after 2006. Panel (b) also shows how much of the picture depends on the unknown category: in 2004–2005 and again in 2007–2009 the identified set is wide, while after 2013 it narrows almost to a point because few deaths were left unclassified.

Count Models of Civilian Harm (H1, H2)

Table 3 reports the NB2 estimates. Overdispersion is overwhelming: the likelihood-ratio test of $\alpha = 0$ gives statistics between 279 and 404 ($p < 0.001$ in every model), and Pearson dispersion in the corresponding Poisson fits ranges from 34 to 76. Poisson inference would therefore understate standard errors several-fold.

In model M1 the trend coefficient is −4.882 (sandwich SE 1.620, $p = 0.003$). On an annual scale this is an incidence-rate ratio of $\exp(-0.488) = 0.614$, so reported civilian deaths per strike fell by roughly 39 per cent a year across the campaign. The result supports the first part of H1. Model M2 adds the PPG indicator. Its coefficient is small and imprecise (0.227, SE 1.811, $p = 0.900$), the trend survives (−5.083, $p = 0.029$), and both AIC and BIC worsen. Once the trend is allowed for, the 2013 guidance adds nothing that the data can detect.

Model M3 adds log strike tempo. Its coefficient is −0.748 (SE 0.342, $p = 0.029$), which implies an elasticity of civilian deaths with respect to strikes of $1 - 0.748 = 0.25$, with a 95 per cent interval of roughly −0.42 to 0.92. Proportional scaling (an elasticity of one) is rejected, which supports H2. Doubling the number of strikes in a year was associated with only about a 19 per cent rise in reported civilian deaths. In this specification the trend loses significance ($p = 0.644$) and the PPG coefficient turns negative but remains insignificant ($p = 0.154$). Trend, tempo and the post-2013 period are strongly collinear in fourteen observations, so the data can establish that civilian harm per strike fell and that it scaled sub-proportionally with tempo, but they cannot apportion the decline cleanly among time, tempo and policy.

Table 3. Negative Binomial (NB2) Models of Annual Civilian Deaths with Strike Exposure

Variable	M1	M2	M3
Constant	0.092 (0.607)	0.025 (0.595)	0.662 (0.662)
Trend (decades)	−4.882*** (1.620)	−5.083** (2.322)	−1.122 (2.432)
Post-2013 PPG regime	—	0.227 (1.811)	−2.525 (1.770)
Log strike tempo (centred)	—	—	−0.748** (0.342)
ln α	0.799** (0.396)	0.798* (0.414)	0.531 (0.487)
Annual IRR of trend	0.614	0.601	0.894
Tempo elasticity (1 + γ)	—	—	0.252
Log-likelihood	−48.34	−48.33	−46.54
AIC / BIC	102.67 / 104.59	104.66 / 107.22	103.08 / 106.28
LR test $\alpha = 0$ ($\frac{1}{2}\chi^2$)	404.40***	403.89***	278.63***
Poisson Pearson dispersion	71.23	76.31	34.05

Note: Dependent variable: civilian deaths; offset ln(strikes). HC1-scaled sandwich standard errors in parentheses. IRR = incidence-rate ratio. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. T = 14.

Gradual Decline or Regime Break? (H1)

Table 4 and Figure 3 report the break tests. The LR sequence peaks at 2007, with a statistic of 5.22, and declines almost monotonically thereafter. The parametric bootstrap puts the 95 per cent critical value at 9.25, and the bootstrap p-value of the sup-LR statistic is 0.281. There is no statistically significant break once a trend is in the model. The BIC-weighted posterior

assigns 0.475 to a break at 2007 and 0.130 to no break at all, with the remaining mass spread thinly over later years; the posterior probability of a break in 2013, the year of the PPG, is 0.035.

Table 4. Structural-Break Tests for the Civilian-Harm Rate

Candidate break year	2007	2008	2009	2010	2011	2012	2013	2014
LR(τ)	5.22	2.42	1.23	1.11	0.08	0.37	0.01	0.20
BIC posterior weight	0.475	0.117	0.065	0.061	0.036	0.042	0.035	0.038
sup-LR [bootstrap p]	5.22 [0.281]							
Bootstrap critical values 90 / 95 / 99%	7.99 / 9.25 / 12.26							
Posterior weight, no break	0.130							

Note: Base model NB2 with offset ln(strikes) and linear trend; each column adds I(t ≥ τ). Bootstrap: 299 parametric replications from the fitted no-break model, re-estimating the full sup-LR sequence in each. Trimming ≈ 15% at each end.

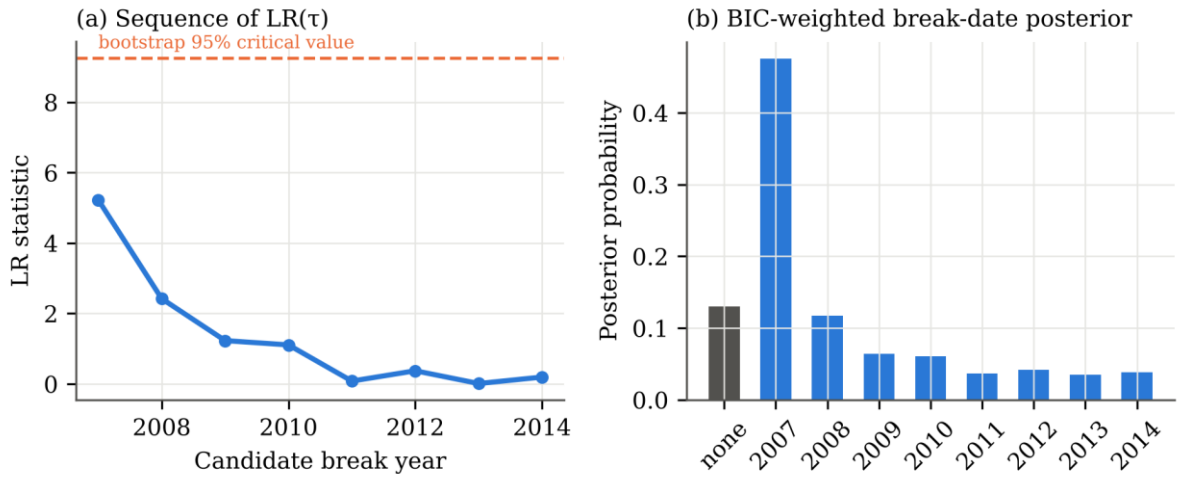


Figure 3. (a) Likelihood-ratio statistics by candidate break year, with the bootstrap 95% critical value; (b) BIC-weighted posterior probabilities of break dates, including the no-break model.

The modal 2007 date is not a technological event. It falls immediately after the 2006 observation, and the break indicator effectively absorbs that single outlier (its IRR is 0.058 with SE 1.118 on the log scale). The substantive reading is that the record shows a steep, gradual fall in reported civilian harm per strike from 2007 onwards, consistent with cumulative improvement in ISR, fusion and targeting procedure, and with changes in what was counted. It does not show a discontinuity that could be credited to the 2013 guidance or to any single system. H1 is supported.

Composition of Deaths

The Dirichlet-multinomial model (Table 5) fits far better than the multinomial logit: the likelihood-ratio statistic is 251.3 for three extra parameters. The median precision is 13.4, which implies an intra-class correlation of 0.077. Deaths cluster within years much more than independent sampling allows, which is what one expects when a handful of strikes produce most casualties.

The average marginal effect of the trend on the expected militant share is +0.417 per decade (SE 0.143, p = 0.003). The corresponding effects on the civilian share (−0.202, SE 0.119, p = 0.090) and the unknown share (−0.216, SE 0.131, p = 0.098) are negative and of similar size. The fitted civilian share falls from 11.7 per cent at 2008 covariate values to 2.6 per cent at 2015 values. One point deserves emphasis. Over time, deaths were reclassified out of the unknown category roughly as fast as they

moved out of the civilian category. Whether that reflects better identification or a more expansive definition of "militant" cannot be determined from the counts themselves.

Table 5. Dirichlet-Multinomial Regression of the Composition of Deaths

Parameter	Militants	Civilians	Unknown
Constant	1.781** (0.872)	−0.297 (0.833)	−0.211 (0.830)
Trend (decades)	1.186 (3.956)	−2.008 (3.842)	−2.070 (3.626)
Log strike tempo	1.019 (0.704)	0.658 (0.551)	0.891 (0.563)
AME of trend on expected share	0.417*** (0.143)	−0.202* (0.119)	−0.216* (0.131)
Log-likelihood: Dirichlet-multinomial / multinomial	−81.15 / −206.82		
LR test of overdispersion	251.35***		
Median precision Σa / intra-class correlation	13.40 / 0.077		

Note: Coefficients are on the log-concentration scale; sandwich standard errors in parentheses. AME = average marginal effect per decade, delta-method standard errors. *** p < 0.01, ** p < 0.05, * p < 0.1.

Partial Identification and Randomization Inference

Table 6 shows what can be said without any assumption about the unknown category. Across 2004–2017 the civilian share lies between 9.3 and 18.2 per cent, and the 90 per cent Imbens–Manski interval for the identified set runs from 2.6 to 26.2 per cent. Splitting the campaign at 2010, the year of peak tempo and the point often cited as the start of a "precision" phase, the identified set falls from [19.4, 34.7] per cent before to [4.3, 10.0] per cent after. The identified set for the change is [−30.4, −9.4] percentage points. Because it excludes zero, the sign of the decline is identified whatever the true status of the unknown dead. Civilian deaths per strike fell from 1.99 to 0.29, and randomization inference on the annual ratios gives p = 0.006.

The same exercise at 2013 gives a smaller identified change, [−19.0, −7.7] points, but randomization inference on per-strike civilian deaths no longer rejects (p = 0.371): the post-2013 years have few strikes and the pre-2013 mean is dominated by 2006. The bounds analysis thus reinforces the count models. The direction of change is robust, while its timing and attribution are not.

Table 6. Manski Bounds on the Civilian Share and Randomization Inference

Comparison	Pre-period bounds	Post-period bounds	Identified set for change	Civilians per strike (pre → post)	RI p-value
Full period 2004–2017	[9.3%, 18.2%]	—	—	0.70	—
Split at 2010	[19.4%, 34.7%]	[4.3%, 10.0%]	[−30.4, −9.4] pp	1.99 → 0.29	0.006
Split at 2013 (PPG)	[10.5%, 20.5%]	[1.5%, 2.7%]	[−19.0, −7.7] pp	0.83 → 0.09	0.371

Note: Lower bound = civilians/total deaths; upper bound = (civilians + unknown)/total deaths. 90% Imbens–Manski interval for the full-period identified set: [2.6%, 26.2%] (2,000 year-level bootstrap replications). RI = randomization inference on the difference in mean annual civilians per strike, 20,000 permutations. pp = percentage points.

How Far Do the Sources Disagree? (H3)

Figure 4 and Table 7 confront the ground-truth problem directly. The four independent Pakistan estimates put the civilian share at 3.0 per cent (Pakistan Ministry of Defence), 9.0 per cent (New America), 19.7 per cent (leaked Pakistani internal assessment) and 21.3 per cent (TBIJ). Their confidence intervals barely overlap. Cochran's Q is 183.9 on three degrees of freedom ($p < 0.001$) and I^2 is 98.4 per cent, so almost all of the variation among sources reflects genuine disagreement in coverage, definitions and method rather than sampling error.

The Paule–Mandel random-effects estimate of the civilian share is 10.4 per cent, with an HKSJ 95 per cent confidence interval of 2.3 to 36.7 per cent. The 95 per cent prediction interval, which describes where a fifth credible source might land, runs from 0.1 to 93.1 per cent. The DerSimonian–Laird estimate is very close (10.5 per cent; interval 1.7 to 44.3 per cent). Adding the DNI's all-theatre figure lowers the pooled share to 8.4 per cent (2.5 to 24.5 per cent) without reducing heterogeneity (I^2 = 98.4 per cent). Leave-one-out estimates range from 8.2 per cent (dropping the leaked assessment) to 15.3 per cent (dropping the Ministry of Defence). Put bluntly, the open record does not pin down the civilian share of deaths in the most studied strike campaign of the century to within an order of magnitude.

Table 7. Random-Effects Meta-Analysis of Competing Civilian-Share Estimates, Pakistan

Source	Period	Civilian share	95% interval	Logit SE
New America	2004–2018	9.0%	[6.5%, 12.5%]	0.186
Bureau of Investigative Journalism	2004–Feb 2020	21.3%	[10.5%, 38.4%]	0.427
Pakistan Ministry of Defence	2008–2013	3.0%	[2.4%, 3.8%]	0.124
Pakistan internal assessment (leaked)	2006–2009	19.7%	[17.0%, 22.7%]	0.092
Pooled, Paule–Mandel + HKSJ (k = 4)		10.4%	[2.3%, 36.7%]	τ^2 = 0.965
Pooled, DerSimonian–Laird + HKSJ		10.5%	[1.7%, 44.3%]	τ^2 = 1.392
95% prediction interval (PM)			[0.1%, 93.1%]	
Heterogeneity: Q (df = 3) / I^2		183.9*** / 98.4%		
Sensitivity: adding US DNI 2016 (k = 5)	2009–2015	8.4%	[2.5%, 24.5%]	I^2 = 98.4%

Note: Analysis on the logit scale; ranges reported by sources treated as 95% intervals. Long War Journal and US-official figures report civilians only and cannot be expressed as shares. HKSJ = Hartung–Knapp–Sidik–Jonkman. *** $p < 0.01$.

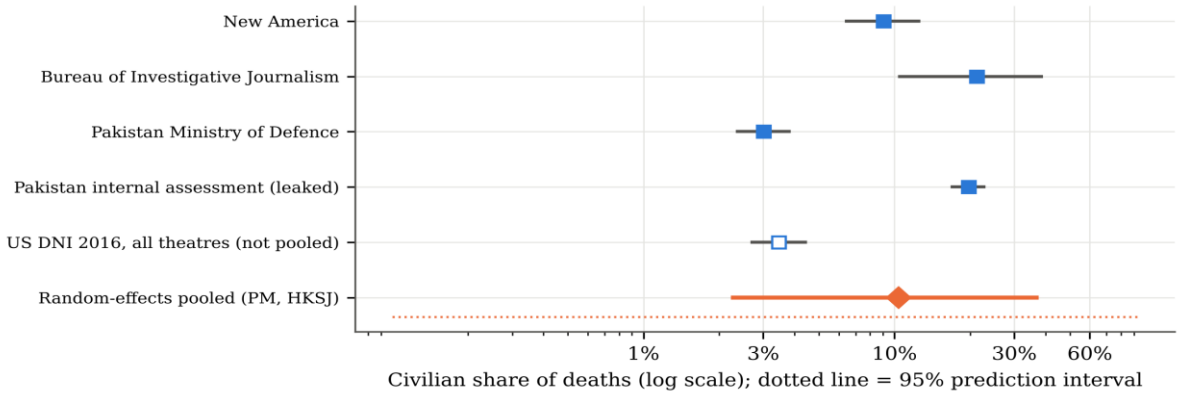


Figure 4. Forest plot of competing civilian-share estimates for US strikes in Pakistan (log scale). The DNI figure (hollow marker) covers all theatres and is excluded from the pooled estimate. The dotted line is the 95% prediction interval.

Cross-Theatre Heterogeneity

Table 8 and Figure 5 propagate TBIJ's reported ranges across the four theatres. Pakistan has the highest expected civilian share (21.7 per cent; 90 per cent interval 13.1 to 32.1 per cent) and is the highest-share theatre in 76.5 per cent of draws. Yemen follows (16.7 per cent; top in 22.2 per cent of draws). Afghanistan (9.1 per cent) and Somalia (4.2 per cent) are lower; Somalia is never the highest. The probability that Pakistan's civilian share exceeds Yemen's is 0.77, so even this ordering is uncertain. Deaths per strike vary by a factor of fourteen, from 0.54 in Afghanistan, where the count includes all US air strikes, to 7.61 in Pakistan. The cross-theatre Spearman correlation between lethality per strike and civilian share averages 0.22 and is positive in 63 per cent of draws: weak evidence that more lethal strikes carried proportionally more civilian harm. Children account for a median 27 per cent of civilian deaths in Pakistan, 24 per cent in Yemen, 21 per cent in Afghanistan and 13 per cent in Somalia.

Table 8. Monte Carlo Propagation of Reported Ranges Across Theatres (100,000 Draws)

Theatre	Min. strikes	Civilian share, mean	90% interval	P(highest share)	Deaths per strike	Children / civilians (median)
Afghanistan	13,072	9.1%	[4.2%, 16.2%]	0.013	0.54	20.7%
Pakistan	430	21.7%	[13.1%, 32.1%]	0.765	7.61	27.2%
Somalia	202	4.2%	[1.2%, 7.2%]	0.000	6.45	12.8%
Yemen	336	16.7%	[13.7%, 20.2%]	0.222	3.58	23.6%

Note: Civilian and total deaths drawn uniformly within TBIJ's reported low–high ranges; totals constrained to be at least civilians. Afghanistan counts include all US air strikes. Mean Spearman ρ (deaths per strike, civilian share) across draws = 0.22; positive in 63% of draws.

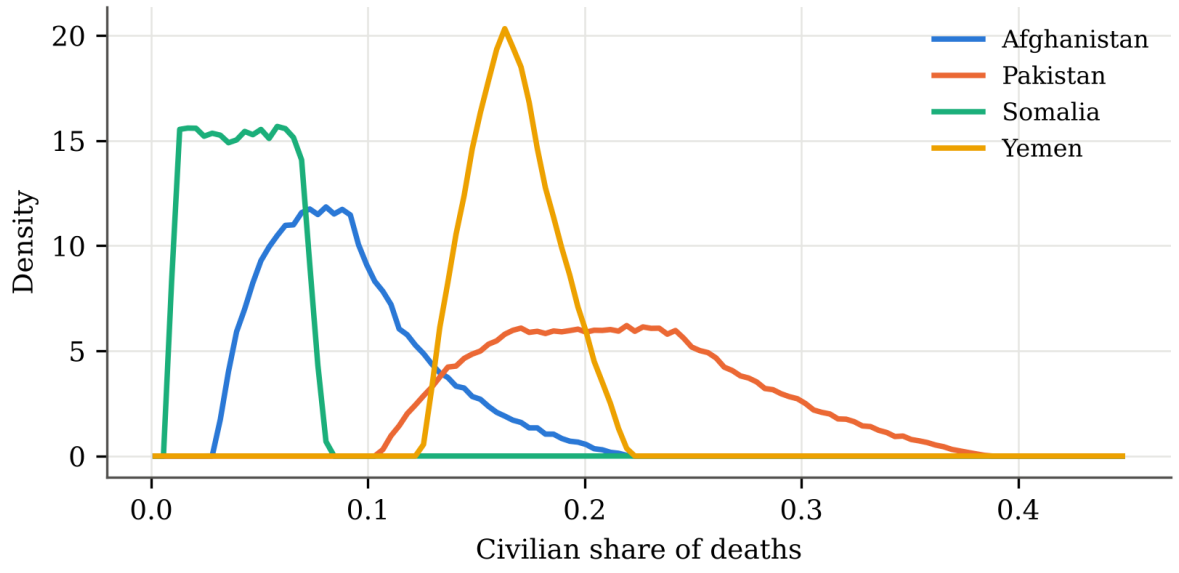


Figure 5. Monte Carlo distributions of the civilian share of deaths by theatre, propagating TBIJ's reported ranges.

The Airwars data for Somalia illustrate the same problem from another angle. Of 267 to 435 civilian deaths alleged in US actions, only 93 to 170 (34.8 to 39.1 per cent) are graded "confirmed or fair"; the rest are weak, contested or discounted. Evidential thresholds, not only events, determine the civilian figure that reaches the public record.

Deep Learning Results (H3, H4)

Table 9 compares out-of-sample performance. In leave-one-year-out validation the deep ensemble records a Poisson deviance of 1,598, a mean absolute error of 44.4 civilian deaths and a root mean squared log error of 2.31. The NB2 model does better on every metric (1,070; 37.0; 1.83), and the Poisson GLM, despite its misspecified variance, gives the most accurate point predictions (851; 26.6; 1.58). All three beat the intercept-only benchmark. The Diebold–Mariano statistic for the ensemble against NB2 is 1.63 (p = 0.128): the deep model is worse, though not significantly so in fourteen folds. Removing 2006 does not change the ranking. H4 is supported: in data of this size and quality, added representational capacity buys variance, not accuracy (Figure 6a).

Table 9. Out-of-Sample Performance: Deep Learning Versus Count-Data Econometrics

Model	Poisson deviance	MAE	MAE excl. 2006	RMSLE
Deep MLP ensemble (3 hidden layers, 10 members)	1,598.3	44.40	40.72	2.312
NB2 GLM (M2)	1,069.6	36.95	32.82	1.835
Poisson GLM (M2)	850.8	26.61	21.72	1.576
Intercept-only NB	2,941.2	111.01	112.47	2.797
Diebold–Mariano, MLP vs NB2 (squared log error)	t = 1.63, p = 0.128			
LSTM walk-forward forecasts of strike tempo, 2011–2017	RMSE (log)	MAE (log)		
LSTM ensemble (8 units, window 3, 5 members)	0.881	0.764		
Random walk	0.660	0.605		
AR(1), least squares	0.857	0.776		

Note: Leave-one-year-out cross-validation (T = 14) for civilian deaths with ln(strikes) offset. MAE = mean absolute error; RMSLE = root mean squared log error. LSTM forecasts are one step ahead, retrained each year on data available up to that year.

The LSTM forecasts of strike tempo tell the same story. Over 2011–2017 its root mean squared log error is 0.881, against 0.857 for AR(1) and 0.660 for a naive random walk. Strike tempo in Pakistan was driven by decisions in Washington and Islamabad, not by a stable temporal pattern a recurrent network could learn from a dozen observations.

The label-noise experiment is the most direct test of H3. Across 24 label sets consistent with the range of sources and five seeds per set, 92.9 per cent of the predictive variance of the ensemble is attributable to label uncertainty and only 7.1 per cent to model uncertainty. The label share exceeds 90 per cent in ten of fourteen years and is lowest (41 per cent) in 2004, the year with a single strike (Figure 6b). No amount of model refinement would remove this variance, because it lies in the targets rather than in the learner.

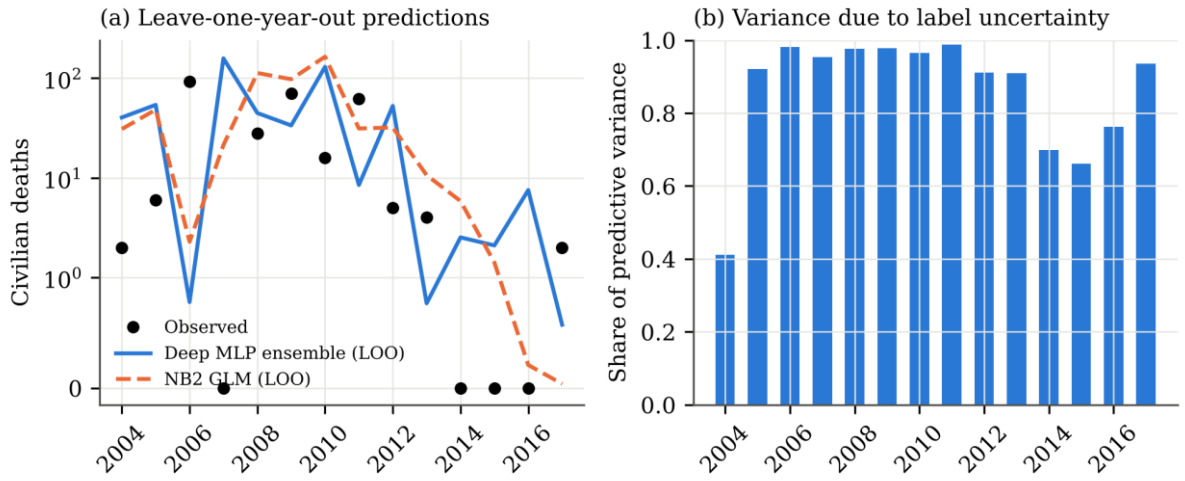


Figure 6. (a) Leave-one-year-out predictions of civilian deaths from the deep ensemble and the NB2 model (symmetric log scale); (b) share of the ensemble's predictive variance attributable to label uncertainty, by year.

The experiment exposes a subtler hazard. For every one of the 120 networks, switching on the PPG indicator lowers predicted civilian deaths, by a factor of 0.28 on average (90 per cent range 0.15 to 0.47). A reader shown only these networks would conclude that the 2013 guidance cut civilian harm by about 70 per cent. The NB2 model, facing the same data, says that the post-2013 effect cannot be separated from the trend ($p = 0.900$ in M2). The networks are not wrong because they agree with each other; they agree because they share the same confounded design, in which the policy indicator is almost a step function of time. Here, agreement within an ensemble measures stability of fit, not identification. For this counterfactual, seed-to-seed variance (0.094 on the log scale) actually exceeds label-to-label variance (0.043), a reminder that confident-looking deep attributions can rest on very little.

Throughput and the Capacity for Human Review (H5)

Table 10 sets the reported throughput figures side by side. The Gospel was described as generating 100 targets a day in 2021, against a pre-AI baseline of about 50 a year in Gaza, an acceleration of 730 times. An NPR estimate of 200 targets in ten to twelve days against 100 targets by twenty officers in 300 days implies a 55-fold increase in daily rate. In the first 35 days of the 2023 war, Israeli figures imply about 429 targets attacked a day. The Maven Smart System was reported at 1,000 recommendations an hour, served by a targeting cell of about 20 people, compared with roughly 2,000 in a comparable cell during the 2003 Iraq war, a hundredfold compression. Reported Maven users grew by a factor of about 3.3 a year between early 2024 and late 2026, a doubling time of roughly seven months, and the Palantir contract ceiling rose by 166 per cent in a year.

Table 10. Reported Throughput of Algorithmic Targeting Systems and Implied Review Capacity

Indicator	Value	Basis
Gospel target-generation acceleration (2021 vs pre-AI baseline)	730×	100/day × 365 vs 50/year
Gospel vs analyst daily rate (NPR estimate)	54.5×	18.2/day vs 0.33/day
Targets attacked per day, first 35 days of 2023 war	428.6	15,000 / 35 days
Lavender individuals flagged / reported accuracy	37,000 / 90%	+972/Local Call sources
Implied misidentified at reported accuracy	3,700	37,000 × (1 − 0.90)

Indicator	Value	Basis
95% interval if accuracy came from validation n = 100 / 500 / 1,000	2,059–6,458 / 2,839–4,790 / 3,069–4,446	Beta posterior
Cumulative review time at 20 seconds per target	205.6 hours	37,000 × 20 s
MSS throughput / cell size / OIF-era cell	1,000 recs/h / 20 / 2,000	NGA Director; secondary reporting
Maximum mean review time sustainable by a 20-person cell	1.2 minutes	20 × 60 / 1,000
Analysts needed (P_wait ≤ 0.2) at 20 s / 1 / 2 / 5 / 10 / 20 min per rec.	9 / 22 / 40 / 94 / 181 / 354	Erlang-C, λ = 1,000/h
Maven active users, annual growth factor / doubling time	3.3× / 6.9 months	Log-linear fit, Mar 2024–Sep 2026

Note: All throughput figures are as reported by officials or journalists; none has been independently verified, and some (flagged in the workbook) come from secondary or market commentary. The Erlang-C results are capacity arithmetic, not a model of any actual organization.

The queueing results make the constraint concrete (Figure 7). If 1,000 recommendations arrive each hour, a 20-person cell working without pause can spend at most 1.2 minutes on each. Keeping the probability of a recommendation waiting below 20 per cent requires 22 analysts at one minute per recommendation, 94 at five minutes, 181 at ten minutes and 354 at twenty. Any review that resembles legal and contextual judgment of an individual target therefore requires either a very different staffing ratio from the one reported or a throughput far below the headline figure. The reported 20-second review per Lavender target sits at the extreme low end of the curve; even so, reviewing all 37,000 flagged individuals once would have taken about 206 analyst-hours. At the reported accuracy, about 3,700 people would have been flagged in error; if the 90 per cent figure came from a validation sample of 100 cases, the 95 per cent interval for that number would run from roughly 2,100 to 6,500. H5 is supported: reported throughput and reported staffing leave little room for substantive human review of each recommendation.

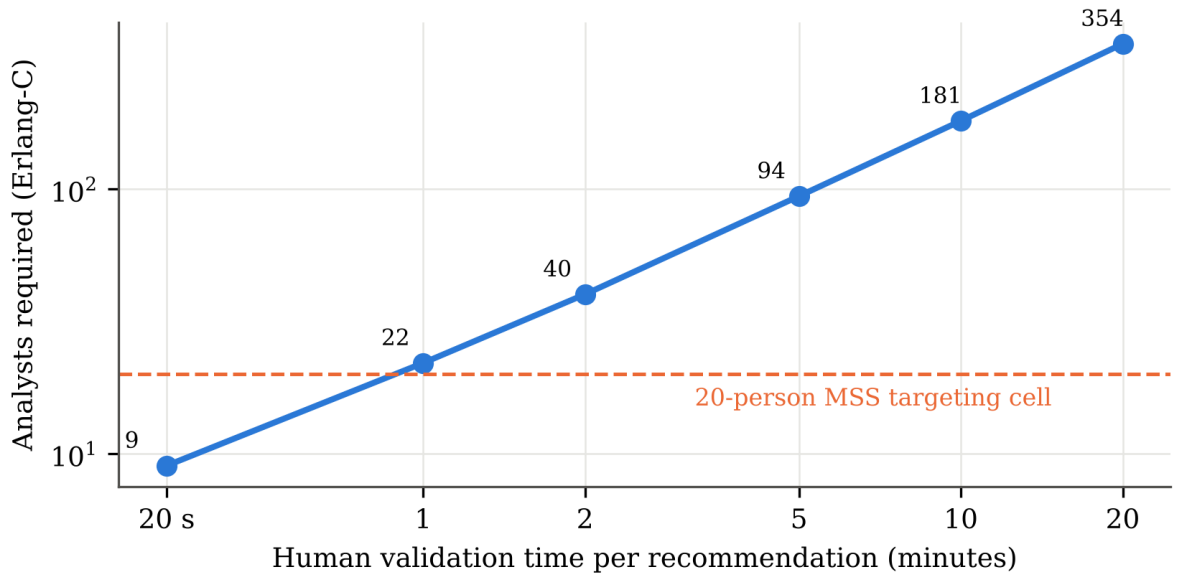


Figure 7. Analysts required to keep the probability of waiting below 0.2 at 1,000 recommendations per hour, by human review time per recommendation (Erlang-C). The dashed line marks the reported 20-person targeting cell.

Table 11. Summary of Hypothesis Tests

Hypothesis	Principal evidence	Verdict
H1: gradual decline, not regime break	NB trend IRR 0.614/yr (p = 0.003); sup-LR p = 0.281; P(break 2013) = 0.035; Manski sign identified	Supported
H2: non-proportional scaling with tempo	Tempo elasticity 0.25 (95% CI −0.42 to 0.92)	Supported
H3: label disagreement dominates	I^2 = 98.4%; PI 0.1–93.1%; 92.9% of predictive variance from labels	Supported
H4: deep models do not beat parsimonious models	LOO deviance MLP 1,598 vs NB2 1,070 vs Poisson 851; LSTM worse than random walk	Supported
H5: throughput exceeds review capacity	1.2 min per recommendation for a 20-person cell; 94 analysts needed at 5 min	Supported

Discussion

The results support a view of algorithmic precision that is more qualified than either its advocates or its critics usually allow. Reported civilian harm per strike in Pakistan really did fall, steeply and on every specification. The sign of the fall survives the most conservative treatment of the unknown dead, and it survives randomization inference that assumes nothing about distributions. The decline is consistent with the claim that persistent ISR, better fusion and improving analytic tools allowed strikes to discriminate more finely as the campaign matured. That much the precision argument can claim.

What it cannot claim is attribution. The decline is gradual. The break tests find no discontinuity once the trend is modelled, and the period most often linked to policy reform, 2013, receives almost no posterior weight. Tempo, time and governance are tangled together in a short series, and the data do not separate them. This matters because the precision argument is frequently made in a stronger, causal form: that a particular system, a particular guidance document or "AI" in general made strikes more discriminating. The open record does not support statements of that form, and the deep-learning experiment shows how easily they can be manufactured. One hundred and twenty networks agreed that the 2013 guidance cut civilian harm by about 70 per cent. A count model with honest standard errors could not detect the effect at all. Ensembles of flexible learners fitted to confounded designs produce unanimous answers to questions the data cannot answer, and that is precisely the pattern that makes black-box outputs persuasive in high-stakes settings (Rudin, 2019).

The sub-proportional relationship between tempo and civilian harm admits several readings, and we do not think the data choose among them. It is consistent with learning by doing, with the selection of easier targets as the programme expanded, and with fewer casualty reports per strike as strikes became routine and media attention moved elsewhere. Each reading has different implications for the "industrialization" critique. What the result does rule out, at least in reported data, is the simplest version of that critique, in which more strikes mean proportionally more civilian dead.

The composition model points to the label layer as the place where precision is partly made. The militant share of deaths rose by about 42 percentage points per decade, and the unknown share fell as fast as the civilian share. Those shifts may reflect better identification. They may also reflect changes in who was counted as a militant; reporting that the United States treated military-age males in strike zones as combatants unless shown otherwise is part of the record that critics cite (Suchman, 2020). Counts alone cannot distinguish these mechanisms. For the study of algorithmic targeting this is not a side issue, because the same assessment conventions that shape public casualty figures shape the labels on which targeting and assessment models are trained. A model trained on labels in which "unknown" drifts into "militant" will learn that drift as signal.

The meta-analysis quantifies the scale of the problem. With I^2 near 98 per cent and a prediction interval running from a fraction of a per cent to more than ninety, the disagreement among credible sources is not noise around a known value. It is structural, rooted in access, definitions and incentives. The label-noise experiment translates this into the language of machine learning: more than nine-tenths of what a model of civilian harm is uncertain about is uncertainty in its training targets. This connects to Goldfarb and Lindsay's (2022) argument that AI raises the value of judgment. Our results suggest a corollary. When labels

are contested, the scarce input is not only judgment at the moment of decision but also judgment about what counts as ground truth, and no improvement in model architecture can supply it.

H4 also has a methodological lesson for security studies. The deep ensemble and the LSTM lost to simple count models and to a random walk. This is not a statement that deep learning is useless for intelligence. Image recognition and signals processing involve millions of examples and comparatively objective labels, conditions under which deep networks excel (LeCun et al., 2015). Strategic outcomes, campaign-level harm and the timing of political decisions are observed rarely and labelled contentiously. Treating results from the first domain as evidence for the second is a category error, and much of the enthusiasm for "AI-driven" threat assessment rests on it.

The throughput analysis turns the normative debate on meaningful human control into an accounting identity. If a system produces a thousand recommendations an hour and twenty people review them, each recommendation gets at most seventy-two seconds of human time. That figure is an upper bound before breaks, handovers, deliberation or escalation. Human factors research suggests that operators under such load drift towards accepting recommendations (Cummings, 2004; Parasuraman & Manzey, 2010). The analysis does not show that review is perfunctory in any particular organization, since we observe neither actual workflows nor the share of recommendations that ever reach a strike decision. It shows that the published throughput and staffing figures cannot both be taken at face value alongside claims of substantive review. Some of the three claims must give.

Finally, the strategic transformation the literature anticipates is visible in inputs more clearly than in outcomes. Budget lines, user counts and contract ceilings are rising quickly, and vendor forecasts put the military AI market at $34–45 billion by the mid-2030s. Outcome data, by contrast, are sparse, contested and slow. The asymmetry is itself strategically important. Organizations will be tempted to judge AI programmes by the metrics that are abundant (targets generated, hours saved, users on boarded) rather than by those that matter and are hard to observe, chief among them discrimination and its strategic consequences in asymmetric conflicts where legitimacy is a resource (Kalyvas, 2006; Lyall & Wilson, 2009).

Policy Implications

Four implications follow, each tied to a specific result.

First, claims of algorithmic precision should be audited against bounded, multi-source ground truth. Reporting a single civilian-casualty figure for a programme invites selective citation. Militaries, oversight bodies and researchers should publish identified sets for key quantities, treating unknown or unattributed deaths explicitly, and should state how their figures compare with independent monitors. The Manski and meta-analytic tools used here are inexpensive and could be adopted in civilian-harm assessment reports such as the US Section 1057 reports.

Second, the label layer should be governed. Because assessment conventions become training data, decisions about how to classify the dead are decisions about how future models will behave. Independent review of battle-damage and casualty-assessment conventions, versioning of the labels used to train targeting and assessment models, and routine evaluation of models against alternative label sets should be part of the legal review that Article 36 of Additional Protocol I and national directives already require (U.S. Department of Defense, 2023a).

Third, meaningful human control should be specified in units of time and staff. Political declarations and directives speak of "appropriate levels of human judgment" (U.S. Department of State, 2023). Our queueing analysis shows that this can be operationalized: for any stated throughput and staffing, the maximum review time per recommendation is a matter of arithmetic. States could commit to reporting minimum review times, ratios of recommendations to qualified reviewers, and the share of recommendations rejected, giving auditors a concrete measure against which to test claims of control.

Fourth, causal claims about AI programmes require designs that can support them. Flexible models fitted to short, confounded series will produce confident attributions. Evaluations of whether a system or a policy changed outcomes should pre-specify their identification strategy, report uncertainty that includes label disagreement, and favour transparent models in high-stakes settings (Rudin, 2019). For states such as Pakistan, which have borne the costs of these campaigns and now face their own decisions about military AI in a South Asian balance where India outspends Pakistan by a factor of 7.7 (Stockholm International Peace Research Institute, 2026), the same discipline applies to domestic procurement and to diplomacy in multilateral forums on autonomous weapons.

Limitations and Future Research

The study has clear limits. The Pakistan series contains fourteen annual observations, which restricts the complexity of any model and the power of every test; we have chosen estimators and inference procedures (sandwich errors, parametric

bootstrap, randomization inference, HKSJ intervals) suited to small samples, but power remains modest. Annual aggregation hides the strike-level variation through which any effect of algorithmic tools would operate, and strike-level data with consistent coding across sources would permit much sharper designs. The trend variable proxies' capability maturation but is not a measurement of it; open-source, time-varying measures of ISR coverage, fusion and model deployment do not exist for this campaign. The cross-theatre comparison uses four aggregates, and Afghanistan's count includes all air strikes. Throughput figures for Gaza and the Maven Smart System are reported claims, some from secondary or market sources, and our queueing analysis is capacity arithmetic rather than a description of any organization. Finally, the label-noise experiment encodes the range of published sources, not the unknown truth; it shows how much the sources' disagreement matters, not which source is right.

Future research could extend the approach to strike-level data for Yemen and Somalia, where Airwars and New America provide event coding; combine casualty sources through formal measurement-error or latent-class models; use document analysis of declassified assessment conventions to model label drift directly; and study human review empirically through controlled experiments with intelligence analysts working at varying recommendation rates.

Conclusion

The idea that deep learning makes war more precise is now embedded in budgets, doctrine and public justification. This article has asked how much of that idea the open record can support. The answer is mixed. Reported civilian harm per strike in the largest covert campaign of the drone era fell steeply and robustly, and civilian deaths grew far less than proportionally with strike tempo. But the decline was gradual rather than tied to any datable change, the sources disagree about its level by more than an order of magnitude, and deep learning applied to the record adds confidence without adding knowledge. Reported throughput for current targeting systems, set against reported staffing, leaves only seconds to a minute or so of human time per recommendation.

Precision, in short, is not a property of algorithms alone. It belongs to the whole architecture, and especially to its weakest links: the labels that define success and the human time available to judge each output. Strategies that rely on algorithmic intelligence in asymmetric war will be only as discriminating, and only as legitimate, as those two links allow. Measuring them honestly is the first step towards governing them.

References

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

Airwars. (2026). U.S. Forces in Somalia [Conflict data]. https://airwars.org/conflicts/us-forces-in-somalia

Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. Survival, 59(5), 117–142. https://doi.org/10.1080/00396338.2017.1375263

Amoore, L. (2009). Algorithmic war: Everyday geographies of the War on Terror. Antipode, 41(1), 49–69. https://doi.org/10.1111/j.1467-8330.2008.00655.x

Amoore, L. (2020). Cloud ethics: Algorithms and the attributes of ourselves and others. Duke University Press.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. Econometrica, 61(4), 821–856. https://doi.org/10.2307/2951764

Arkin, R. C. (2009). Governing lethal behavior in autonomous robots. CRC Press.

Arreguín-Toft, I. (2001). How the weak win wars: A theory of asymmetric conflict. International Security, 26(1), 93–128. https://doi.org/10.1162/016228801753212868

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. Econometrica, 66(1), 47–78. https://doi.org/10.2307/2998540

Berman, E., Felter, J. H., & Shapiro, J. N. (2018). Small wars, big data: The information revolution in modern conflict. Princeton University Press.

Berman, E., Shapiro, J. N., & Felter, J. H. (2011). Can hearts and minds be bought? The economics of counterinsurgency in Iraq. Journal of Political Economy, 119(4), 766–819. https://doi.org/10.1086/661983

Biddle, S. (2004). Military power: Explaining victory and defeat in modern battle. Princeton University Press.

Bode, I., & Huelss, H. (2018). Autonomous weapons systems and changing norms in international relations. Review of International Studies, 44(3), 393–413. https://doi.org/10.1017/S0260210517000614

Brose, C. (2020). The kill chain: Defending America in the future of high-tech warfare. Hachette Books.

Buchanan, B., & Imbrie, A. (2022). The new fire: War, peace, and democracy in the age of AI. MIT Press.

Bureau of Investigative Journalism. (2020). Drone warfare [Database]. https://www.thebureauinvestigates.com/projects/drone-war

Byman, D. (2013). Why drones work: The case for Washington's weapon of choice. Foreign Affairs, 92(4), 32–43.

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (2nd ed.). Cambridge University Press.

Cederman, L.-E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? Science, 355(6324), 474–476. https://doi.org/10.1126/science.aal4483

Chamayou, G. (2015). A theory of the drone (J. Lloyd, Trans.). The New Press.

Cohen, E. A. (1996). A revolution in warfare. Foreign Affairs, 75(2), 37–54.

Cronin, A. K. (2013). Why drones fail: When tactics drive strategy. Foreign Affairs, 92(4), 44–54.

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In AIAA 1st Intelligent Systems Technical Conference (AIAA 2004-6313). American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313

Davies, H., McKernan, B., & Sabbagh, D. (2023, December 1). 'The Gospel': How Israel uses AI to select bombing targets in Gaza. The Guardian.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Faris, L. (2025). Algorithmic targeting in the Iranian–Israeli confrontation: Technical realities, legal thresholds, and the boundaries of human control. F1000Research. https://doi.org/10.12688/f1000research.169794.1

Fearon, J. D., & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. American Political Science Review, 97(1), 75–90. https://doi.org/10.1017/S0003055403000534

Fisher, R. A. (1935). The design of experiments. Oliver & Boyd.

Ford, M., & Hoskins, A. (2022). Radical war: Data, attention and control in the twenty-first century. Hurst.

Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

Garcia, D. (2018). Lethal artificial intelligence and change: The future of international peace and security. International Studies Review, 20(2), 334–341. https://doi.org/10.1093/isr/viy029

Gilli, A., & Gilli, M. (2019). Why China has not caught up yet: Military-technological superiority and the limits of imitation, reverse engineering, and cyber espionage. International Security, 43(3), 141–189. https://doi.org/10.1162/isec_a_00337

Goldfarb, A., & Lindsay, J. R. (2022). Prediction and judgment: Why artificial intelligence increases the importance of humans in war. International Security, 46(3), 7–50. https://doi.org/10.1162/isec_a_00425

Goldman, E. O., & Grossman, R. L. (2026). From data to victory: The race for analytic superiority in warfare. National Defense University Press. https://ndupress.ndu.edu/Media/News/News-Article-View/Article/4560509/

González, R. J. (2022). War virtually: The quest to automate conflict, militarize data, and predict the future. University of California Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of the 3rd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Gregory, D. (2011). From a view to a kill: Drones and late modern war. Theory, Culture & Society, 28(7–8), 188–215. https://doi.org/10.1177/0263276411423027

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520.

Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). Fundamentals of queueing theory (4th ed.). Wiley.

Haas, M. C., & Fischer, S.-C. (2017). The evolution of targeted killing practices: Autonomous weapons, future conflict, and the international order. Contemporary Security Policy, 38(2), 281–306. https://doi.org/10.1080/13523260.2017.1336407

Hambling, D. (2015). Swarm troopers: How small drones will conquer the world. Archangel Ink.

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. Statistics in Medicine, 20(24), 3875–3889. https://doi.org/10.1002/sim.1009

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents–R&D relationship. Econometrica, 52(4), 909–938. https://doi.org/10.2307/1911191

Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., Hoyles, F., Hultman, L., Högbladh, S., Jansen, R., Mouhleb, N., Muhammad, S. A., Nilsson, D., Nygård, H. M., Olafsdottir, G., Petrova, K., Randahl, D., Rød, E. G., Schneider, G., … Vestby, J. (2019). ViEWS: A political violence early-warning system. Journal of Peace Research, 56(2), 155–174. https://doi.org/10.1177/0022343319823860

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horowitz, M. C. (2010). The diffusion of military power: Causes and consequences for international politics. Princeton University Press.

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. Texas National Security Review, 1(3), 36–57. https://doi.org/10.15781/T2639KP49

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. Journal of Strategic Studies, 42(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. Econometrica, 72(6), 1845–1857. https://doi.org/10.1111/j.1468-0262.2004.00555.x

International Committee of the Red Cross. (2021). ICRC position on autonomous weapon systems. ICRC.

Jaeger, D. A., & Siddique, Z. (2018). Are drone strikes effective in Afghanistan and Pakistan? On the dynamics of violence between the United States and the Taliban. CESifo Economic Studies, 64(4), 667–697.

Johnson, J. (2021). Artificial intelligence and the future of warfare: The USA, China, and strategic stability. Manchester University Press.

Johnston, P. B. (2012). Does decapitation work? Assessing the effectiveness of leadership targeting in counterinsurgency campaigns. International Security, 36(4), 47–79. https://doi.org/10.1162/ISEC_a_00076

Johnston, P. B., & Sarbahi, A. K. (2016). The impact of US drone strikes on terrorism in Pakistan. International Studies Quarterly, 60(2), 203–219. https://doi.org/10.1093/isq/sqv004

Jordan, J. (2009). When heads roll: Assessing the effectiveness of leadership decapitation. Security Studies, 18(4), 719–755. https://doi.org/10.1080/09636410903369068

Jordan, J. (2014). Attacking the leader, missing the mark: Why terrorist groups survive decapitation strikes. International Security, 38(4), 7–38. https://doi.org/10.1162/ISEC_a_00157

Kalyvas, S. N. (2006). The logic of violence in civil war. Cambridge University Press.

Kania, E. B. (2017). Battlefield singularity: Artificial intelligence, military revolution, and China's future military power. Center for a New American Security.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

King, A. (2024). Digital targeting: Artificial intelligence, data, and military intelligence. Journal of Global Security Studies, 9(2), ogae009. https://doi.org/10.1093/jogss/ogae009

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

Kissinger, H. A., Schmidt, E., & Huttenlocher, D. (2021). The age of AI: And our human future. Little, Brown and Company.

Konaev, M., & Chahal, H. (2021, February 18). Building trust in human-machine teams. Brookings TechStream. https://www.brookings.edu/techstream/building-trust-in-human-machine-teams/

Krepinevich, A. F. (1994). Cavalry to computer: The pattern of military revolutions. The National Interest, 37, 30–42.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lyall, J., & Wilson, I. (2009). Rage against the machines: Explaining outcomes in counterinsurgency wars. International Organization, 63(1), 67–106. https://doi.org/10.1017/S0020818309090031

Mack, A. (1975). Why big nations lose small wars: The politics of asymmetric conflict. World Politics, 27(2), 175–200. https://doi.org/10.2307/2009880

Mahnken, T. G. (2008). Technology and the American way of war since 1945. Columbia University Press.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Manski, C. F. (2003). Partial identification of probability distributions. Springer.

Marshall, A. W. (1993, August 23). Memorandum for the record: Some thoughts on military revolutions—Second version. Office of the Secretary of Defense, Office of Net Assessment.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika, 49(1/2), 65–82. https://doi.org/10.2307/2333468

Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review, 112(2), 358–375. https://doi.org/10.1017/S0003055417000570

New America. (2018). The drone war in Pakistan [Database]. New America International Security Program.

Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 70, 1373–1411. https://doi.org/10.1613/jair.1.12125

Office of the Director of National Intelligence. (2016). Summary of information regarding U.S. counterterrorism strikes outside areas of active hostilities. ODNI.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. Journal of Research of the National Bureau of Standards, 87(5), 377–385. https://doi.org/10.6028/jres.087.022

Payne, K. (2018). Artificial intelligence: A revolution in strategic affairs? Survival, 60(5), 7–32. https://doi.org/10.1080/00396338.2018.1518374

Payne, K. (2021). I, warbot: The dawn of artificially intelligent conflict. Hurst.

Price, B. C. (2012). Targeting top terrorists: How leadership decapitation contributes to counterterrorism. International Security, 36(4), 9–46. https://doi.org/10.1162/ISEC_a_00075

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. Journal of the American Statistical Association, 55(290), 324–330. https://doi.org/10.1080/01621459.1960.10482067

Roff, H. M. (2014). The strategic robot problem: Lethal autonomous weapons in war. Journal of Military Ethics, 13(3), 211–227.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, 15. https://doi.org/10.3389/frobt.2018.00015

Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. W. W. Norton.

Scharre, P. (2023). Four battlegrounds: Power in the age of artificial intelligence. W. W. Norton.

Schwarz, E. (2018). Death machines: The ethics of violent technologies. Manchester University Press.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Stockholm International Peace Research Institute. (2026). Trends in world military expenditure, 2025 (SIPRI Fact Sheet). SIPRI. https://www.sipri.org/sites/default/files/Military%20Expenditure%202025.pdf

Suchman, L. (2020). Algorithmic warfare and the reinvention of accuracy. Critical Studies on Security, 8(2), 175–187. https://doi.org/10.1080/21624887.2020.1760587

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1312.6199

U.S. Department of Defense. (2023a). DoD Directive 3000.09: Autonomy in weapon systems. Office of the Under Secretary of Defense for Policy.

U.S. Department of Defense. (2023b). Data, analytics, and artificial intelligence adoption strategy. Chief Digital and Artificial Intelligence Office.

U.S. Department of State. (2023). Political declaration on responsible military use of artificial intelligence and autonomy. Bureau of Arms Control, Deterrence and Stability.

United Nations General Assembly. (2023). Lethal autonomous weapons systems (A/RES/78/241). United Nations.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Vincent, B. (2022, October 7). Pentagon remains mum on Project Maven as transition is held up by continuing resolution. DefenseScoop. https://defensescoop.com/2022/10/07/pentagon-remains-mum-on-project-maven-as-transition-is-held-up-by-continuing-resolution/

Vlad, R. O., & Hardy, J. (2025). Signature strikes and the ethics of targeted killing. International Journal of Intelligence and CounterIntelligence. https://doi.org/10.1080/08850607.2024.2382029

Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. Journal of Peace Research, 47(4), 363–375. https://doi.org/10.1177/0022343309356491

Watts, B. D. (2004). Clausewitzian friction and future war (Rev. ed., McNair Paper 68). Institute for National Strategic Studies, National Defense University.

Appendix A. Sources compiled in the study workbook

All URLs accessed 31 October 2026. Figures taken from Wikipedia pages are traced there to the underlying New America and TBIJ datasets.

Table A1. Sources of the Study Workbook

No.	Source	Workbook sheet
1	New America data via Wikipedia, 'American drone strikes in Pakistan'	Pakistan_Yearly
2	Wikipedia, 'Civilian casualties from the United States drone strikes' (TBIJ & New America tables)	Campaigns_TBIJ, Pakistan_Estimates
3	Wikiwand mirror — Pakistan civilian estimates (LWJ, New America, TBIJ)	Pakistan_Estimates
4	Fox News/AP — Pakistan lowers estimate of civilians killed by drone strikes (2013)	Pakistan_Estimates
5	Airwars (TBIJ archive) — Pakistan internal assessment, 75 strikes 2006-09	Pakistan_Estimates
6	TBIJ — Pakistan: Reported US strikes 2018	Pakistan_Yearly note
7	Airwars — U.S. Forces in Somalia (conflict page)	Somalia_Airwars
8	Airwars — US Forces in Somalia (2025 snapshot)	Somalia_Airwars
9	Airwars — US 1057 report 2025 analysis	Somalia_Airwars
10	DefenseScoop — Pentagon AI budget request FY2025	DoD_AI_Budget
11	DefenseScoop — Pentagon requesting >$3B for AI, JADC2 (FY2024)	DoD_AI_Budget
12	MeriTalk — FY2025 DoD budget request	DoD_AI_Budget
13	CDO Magazine — Pentagon seeks $13.4bn for AI and autonomy FY2026	DoD_AI_Budget

No.	Source	Workbook sheet
14	GrantedAI — Pentagon AI budget beyond DARPA (CDAO, DIU figures)	DoD_AI_Budget
15	Defence Blog — $4.2B Sovereign AI Infrastructure request FY2027	DoD_AI_Budget
16	DefenseScoop — DoD raises Palantir Maven contract to >$1B	Maven_Timeline
17	MeriTalk — NGA expands Maven AI system	Maven_Timeline
18	Tom's Hardware — Pentagon formalizes Maven as core system	Maven_Timeline
19	Nasdaq/MarketBeat — How Maven turns Palantir's biggest risk into strength	Maven_Timeline
20	BD Emerson — Maven Smart System timeline	Maven_Timeline
21	Wikipedia — AI-assisted targeting in the Gaza Strip	AI_Targeting_Gaza
22	The Guardian — 'The Gospel': how Israel uses AI to select bombing targets (2 Dec 2023)	AI_Targeting_Gaza
23	Arab News — Israeli AI used to identify 37,000 targets (citing +972)	AI_Targeting_Gaza
24	SIPRI — Military Expenditure 2025 fact sheet	SIPRI_MilEx_2025
25	SIPRI — press release 27 Apr 2026	SIPRI_MilEx_2025
26	Dawn — Pakistan's defence spending rises 11pc	SIPRI_MilEx_2025
27	Jerusalem Post — SIPRI 2025 (NATO total)	SIPRI_MilEx_2025
28	Visiongain — Military AI Systems Market Report 2026-2036	AI_Market_Forecasts
29	Persistence Market Research — AI in Military Market 2026-2033	AI_Market_Forecasts
30	Diginomica — SIPRI mapping the military AI industry	DoD_AI_Budget note

Note: Full URLs are listed in the 'Sources' sheet of the supplementary workbook.

Appendix B. Estimation settings and replication

Table B1. Estimation Settings

Component	Setting
Count models	NB2 by maximum likelihood (BFGS then Nelder–Mead, three starting values for ln α); HC1-scaled sandwich covariance from numerical scores and Hessian
Break test	Candidate years 2007–2014; 299 parametric bootstrap replications from the fitted no-break NB2; BIC weights over nine models

Component	Setting
Dirichlet-multinomial	Covariates: constant, trend, centred log tempo; nine parameters; sandwich covariance; AMEs by central differences, delta method
Partial identification	Imbens–Manski interval with critical value 1.645; 2,000 year-level bootstrap replications
Randomization inference	20,000 permutations of year-to-period assignment
Meta-analysis	Logit scale; DerSimonian–Laird and Paule–Mandel τ^2; HKSJ with t(k − 1); prediction interval with t(k − 2)
Monte Carlo	100,000 uniform draws within reported ranges per theatre
Deep MLP	Inputs standardized (trend, PPG, log tempo); layers 16–16–8, tanh; dropout 0.1; L2 10^{-3}; Adam (lr 0.005); 3,000 full-batch epochs; Poisson loss with offset; 10-member ensemble
LSTM	One layer, 8 units, window 3, inputs standardized on training data only; Adam (lr 0.01), 600 epochs, gradient clipping at 1; 5-member ensemble; walk-forward 2011–2017
Label-noise design	24 label sets × 5 seeds; common factor log-uniform on [50/288, 969/288]; year-level lognormal noise with σ = 0.21
Queueing	Erlang-C via Erlang-B recursion; λ = 1,000 per hour; P_wait ≤ 0.2
Software and seeds	Python 3.13.18, NumPy 2.5, SciPy 1.18, pandas 3.0, scikit-learn 1.9, Matplotlib 3.11; master seed 20261031

Note: All neural networks are implemented directly in NumPy (forward pass, back-propagation and Adam), so results do not depend on any deep-learning framework.

2026 by the authors; Journal of *Kashmir Journal of Academic Research and Development*. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).