MEASURING THE SEXUAL-OFFENCE JUSTICE GAP, DIAGNOSING ITS STRUCTURE, AND SIMULATING ITS REPAIR — FROM TWO MILLION RECORDS, WITH TOOLS BORROWED FROM OTHER SCIENCES

Three in a Hundred

Daniel Fat

Inference Institute — Safeguarding

June 2026 • v6.0 •

Citable via its inference.institute URL; a ready-to-copy BibTeX entry ships on the paper page.

Abstract

Of every hundred rapes recorded by the police in England and Wales, about three end in a charge and roughly two in a conviction. This paper takes that number apart and asks what would move it, reading the public record of how the system counts, resolves and fails these cases not as prose but as structure. The problem it addresses is threefold: the numbers that frame the field are treated as if they measured harm when they largely measure the regime that produced them; the outcome of the more than one hundred thousand offences recorded each year is reported as a single national average that hides who the system sheds and where; and the lessons inside hundreds of statutory reviews are extracted by hand, one document at a time, so that nobody can say whether failure has a shape. The gap is the absence of a method that reads all three at the level of structure and turns that structure into something a reformer can act on. Our method borrows formulas from economics, data mining, signal processing, stochastic processes, information theory and graph control theory and applies each, under an explicit structural analogy, to three bodies of evidence: European recorded-crime statistics; roughly two million rows of Home Office police-outcomes open data, on which we fit statistical models of the system's response; and a coded, anonymised corpus of public reviews built with a fully-local language model. The research design pairs descriptive secondary analysis with predictive and explanatory modelling and a design-science extraction pipeline, and the synthesis of the resulting evidence is the analytic core. The findings are concrete. Recorded rape is distributed across European states with a Gini coefficient of 0.52 — as unequally as household income in Brazil — so the counts rank regimes, not danger. At national scale only 3.5% of recorded rapes lead to a charge while half are closed because the victim withdraws support, the charge rate varies almost fourfold across police forces with eighteen of forty-four falling outside statistical control limits, and a calibrated model shows the offence type and the force — the location — are the dominant drivers of whether a case is charged. And the failures inside the reviews are not formless: they arrive in predictable bundles, collapse onto three latent factors dominated by institutional self-protection, follow a decision grammar in which a reviewed case is as likely to end in a retrospective external review as in a prosecution, and sit inside an agency network with no single point of failure — so the breakdown is unread signal, not missing connection. Finally, the simulations make the repair quantitative: supporting complainants so fewer withdraw is the single largest lever, roughly tripling convictions per recorded rape on a stated prior; bringing laggard forces up to the median-to-top charging rate would add one to four thousand charges a year; and reweighting the decision grammar toward early recognition lifts the simulated share of cases reaching prosecution from 39% to 50%. The contribution is a reusable method, a taxonomy of measurement regimes, a set of calibrated outcome models, three reform simulations, a structural leverage index, and an empirical map from each pattern to a named reform owner. The implications are immediate for policing, law and government, and are stated as testable hypotheses; throughout we measure honestly how far a 1.2-billion-parameter local model can be trusted, and we exclude by design any judgement of a complainant's credibility.

Keywords: rape; sexual offences; crime measurement; case outcomes; attrition; charge rate; victim withdrawal; logistic regression; gradient boosting; funnel plot; Monte-Carlo simulation; microsimulation; counterfactual; statutory reviews; safeguarding; network analysis; absorbing Markov chains; association rules; non-negative matrix factorisation; mutual information; graph control theory; local LLM extraction; reproducibility; record integrity; reform

CONTENTS

1 INTRODUCTION

Rape and other sexual offences sit where two systems meet, and both run on inference under uncertainty. A justice system must decide whom to charge and convict on incomplete evidence. A safeguarding system must decide whom to protect before harm recurs. Both lean on the public record: prevalence estimates and recorded-crime counts on one side, and on the other the statutory reviews and inquiries written each time an institution fails badly enough to compel one. That record carries real weight. It sets priorities, frames every public argument about whether things are improving, and is quoted as evidence in Parliament and the press. Yet it is produced by machinery that is rarely examined when the figures or the lessons are repeated. A recorded offence is not counted the same way in two jurisdictions; a survey estimate is a different kind of object from a police count; and the learning inside hundreds of reviews is distilled by hand, one review at a time, by analysts who cannot read them all and cannot show their working.

This is well-charted territory along each separate axis. There is a deep literature on prevalence and on the steep attrition between offences committed and offences punished (Office for National Statistics 2021; Her Majesty's Inspectorate of Constabulary 2014); a contested but mature evidence base on false-allegation rates (Lisak et al. 2010; Kelly et al. 2005); and, quite separately, decades of thematic synthesis of serious case reviews (Brandon et al. 2009; Sidebotham et al. 2016). What is missing — the niche this work occupies — is anything that joins the two halves and reads the failure record the way an analyst reads a dataset rather than a story. The gap is not a shortage of narratives; it is the absence of structure. Nobody has shown, on the public corpus, that institutional failure recurs in measurable combinations, that it has a sequential grammar, or that the points where reform would bite can be located rather than asserted.

This paper occupies that niche with a single, deliberately interdisciplinary move: we treat the record as data and read it with formulas borrowed from sciences that have already solved structurally similar problems. Economists know how to measure how unequally a quantity is spread; retailers know how to find which products predict each other; signal engineers know how to decompose a tangle of signals into a few latent factors; the theory of stochastic processes knows how to describe where a system is absorbed and how long it takes to get there; information theory knows how to quantify how much one variable tells you about another; and graph theorists know how to find the structural points of leverage in a network. None of these tools has been brought to the safeguarding record. Each is licensed here by an explicit analogy, applied to a coded corpus, and read back into the language of reform.

The thesis is simple and, we think, important: *institutional failure to protect people from sexual offences is not random or idiosyncratic — it has a measurable structure and a grammar, and the right cross-domain tools turn that structure into a map for reform.* The argument runs in four movements that follow an offence through the system. First, the numbers that count it mislead, and they do so predictably: an inequality lens shows that recorded counts measure regime, not harm. Second, the outcome it meets can be modelled at national scale: on two million rows of police-outcomes open data, statistical models show that only a small fraction of recorded sexual offences are charged, that the system sheds most of them through victim withdrawal, and that the probability of a charge depends as much on which force area an offence is reported in as on the offence itself — a quantified postcode lottery. Third, the mechanism behind that outcome has a grammar, recovered from the review record: failure

comes in predictable bundles, collapses onto a few latent archetypes, follows a decision pathway with a definite shape, and exposes specific structural leverage points. Fourth, the repair can be *simulated* rather than asserted: Monte-Carlo and counterfactual experiments on the same data put numbers on what each reform would buy — how many more convictions complainant support would yield, how many more charges closing the postcode lottery would add, and how far moving recognition upstream would shift cases toward prosecution. The first three movements describe how the system works now and why it fails; the fourth turns the diagnosis into a quantified, testable account of how it could work better. Our contribution is the method that makes this reading reproducible, the empirical structures, calibrated models and reform simulations it produces, and the map that ties them to the institutions that could act. Two commitments hold throughout, because the subject demands them: the work performs record-integrity analysis and never scores a complainant's credibility, and every claim about an automated pipeline is accompanied by an honest measurement of how far that pipeline can actually be trusted.

2 BACKGROUND AND CONTEXT

The scale of the harm is not in serious dispute, even though its measurement is. Population surveys in England and Wales consistently estimate that a large minority of women and a smaller but substantial fraction of men experience sexual violence in their lifetime, that fewer than one in six victims report to the police, and that of offences that are recorded only a small percentage — of the order of three in a hundred for rape — result in a charge (Office for National Statistics 2021; Her Majesty's Inspectorate of Constabulary 2014). The distance between these figures is the attrition problem, and it means that any single number describing "sexual offending" is a measurement taken at one point in a long and leaky pipeline.

When prevention fails catastrophically, the state commissions a review. In England and Wales these come in several statutory species — serious case reviews and their successor child safeguarding practice reviews, domestic homicide reviews, the thematic reports of the Independent Inquiry into Child Sexual Abuse (IICSA), and standalone public inquiries such as those into Rotherham and Telford (Jay 2014; Crowther 2022; Independent Inquiry into Child Sexual Abuse 2022). Scotland and Northern Ireland run their own equivalents under different law, which is one reason the record resists naive pooling. Each review is written to a broadly common purpose — to establish what happened, identify what went wrong, and recommend change — but in an idiosyncratic voice and structure, and the corpus as a whole is large, growing, and almost never read in aggregate.

The policy reflex that this record has produced is worth naming at the outset, because our findings press on it. Review after review concludes that agencies "failed to share information" or "worked in silos," and the standing response has been to build more information-sharing machinery: shared databases, multi-agency safeguarding hubs, new duties to cooperate. We will show, structurally, that the agencies in these cases were already densely connected — that the network had no missing link to add — and that the failure lay not in the absence of connection but in connection that carried no signal. That distinction, invisible to a narrative reading, is exactly what a structural reading is for.

3 LITERATURE REVIEW

Five bodies of work meet in this paper, and our synthesis is an attempt to make them speak to one another rather than past one another.

Measurement and attrition. The criminological literature has long warned that recorded-crime counts are administrative artefacts as much as measurements of behaviour (Office for National Statistics 2021). Definitions of rape differ in whether they are penetration-specific or consent-based; counting rules differ in whether they tally victims, perpetrators or incidents; and recording practice differs with police reform, public confidence, and the salience of a recent scandal. The English "crime harm" tradition and the Eurostat and UNODC efforts to standardise offence categories under the International Classification of Crime for Statistical Purposes (ICCS) are partial correctives (Eurostat 2026; United Nations Office on Drugs and Crime 2024), but the residual incomparability is severe, and it is rarely quantified rather than merely cautioned against.

Outcomes and the attrition surface. A second literature studies what happens to offences once recorded: the steep attrition from report to charge to conviction, the central and rising role of victim withdrawal, and the recent collapse and partial recovery of charge rates that prompted the Operation Soteria reforms (Her Majesty's Inspectorate of Constabulary 2014; Home Office 2023; Home Office 2025). This work is usually conducted on national aggregates or bespoke case-file samples; the Home Office now publishes the underlying force-by-offence-by-outcome counts as open data (Home Office 2025; Home Office 2026), but they are, to our knowledge, rarely modelled to ask how much of the variation in who gets charged is attributable to the offence and how much to the place. That modelling is one of our contributions.

Synthesis of the review record. The third literature is the thematic synthesis of serious case reviews (Brandon et al. 2009; Sidebotham et al. 2016). This work is invaluable and is the direct ancestor of ours, but it is conducted manually: expert teams read a sample of reviews and report recurring themes. The synthesis is therefore unauditable in the technical sense — the coding cannot be re-run, the sample cannot be expanded without re-reading, and the recurring themes are stated as qualitative findings rather than measured structures. Our debt to this tradition is conceptual; our departure from it is methodological.

Computational and network criminology. A fourth, newer literature applies network and computational methods to crime and safeguarding data (Morselli 2009). Most of it analyses offender or co-offending networks; very little treats the *institutional* response as a network, and we are not aware of prior work that represents a safeguarding case as a decision-pathway graph and compares such graphs across a corpus. The methodological tools we borrow are individually standard — the Gini coefficient (Gini 1912), association-rule mining (Agrawal et al. 1993), non-negative matrix factorisation (Lee and Seung 1999), absorbing Markov chains (Kemeny and Snell 1976), mutual information (Shannon 1948), and k-core and betweenness analysis (Seidman 1983; Freeman 1977) — but their joint application to the safeguarding record is, to our knowledge, new.

Large language models for structured extraction. The fifth literature concerns the use of language models to convert documents into structured data (Ziems et al. 2024). The honest finding in that work is that capability is uneven and that validation against human coding is indispensable. We adopt its discipline directly: a fully local model does the reading, but every category it produces is checked against a human-coded gold standard, and we report where it succeeds and where it fails rather than where we wish it would.

The synthesis across these five is the observation that drives the paper. Measurement scholars have the right caution but no structural method; outcomes scholars have a rich open dataset that is seldom modelled for the relative weight of offence and place; review synthesists have the right material but no reproducible instrument; network criminologists have the instrument but have not pointed it at the institutional record; and the LLM literature supplies the means to build the corpus at scale, provided its reliability is measured rather than assumed. Joining them is the contribution this work sets out to make.

4 CONCEPTUAL AND THEORETICAL FRAMEWORK

The framework rests on one stance and one principle.

The stance is that the system's response to a recorded sexual offence can be represented in three compatible ways: as a *probabilistic outcome*, as a *stochastic process*, and as a *network*. As an outcome, the response to a recorded offence is a draw from a distribution over charge, out-of-court disposal, victim withdrawal, no-suspect and the rest, conditioned on the offence and the place; estimating that distribution is a modelling problem with a right answer that can be checked against held-out data. As a process, a case is a sequence of institutional decisions — a referral, an investigation, a strategy meeting, a plan, a closure, sometimes a prosecution, often an external review — and the regularities in how one decision follows another are the "grammar" of the system. As a network, a case is a set of agencies and decisions joined by their co-occurrence, and the shape of that network — what sits at its centre, what holds it together, where it would break — is the structure of the response. The same case is all three; each representation answers a different question, at a different scale, on a different body of evidence.

The principle is that each analytic tool we import is *licensed by a structural analogy*, not by metaphor. We make the analogies explicit so they can be challenged (Table 1). Measuring how unequally recorded rape is spread across states is formally the same problem as measuring how unequally income is spread across households, so the Lorenz curve and Gini coefficient transfer exactly. Asking which failures co-occur more than chance is the market-basket problem, so association-rule lift transfers exactly. Asking whether many correlated failures are generated by a few hidden causes is the source-separation problem, so non-negative matrix factorisation transfers. Asking whether a force's charge rate is an outlier or just noisy is the institutional-comparison problem that medicine solved with the funnel plot, so its exact binomial control limits transfer. Asking how much of the variation in who gets charged is attributable to the place rather than the offence is a variance-decomposition problem, answered by the deviance accounting of nested generalised linear models and, for the spread itself, by an inequality index. Asking where a case ends up and how long it takes is the absorbing-chain problem. Asking how much a present risk tells you about a coming failure is the channel-capacity problem. Asking which node's removal would most change the structure is the network-control problem. In every case the analogy is structural identity, not

Borrowed tool	Home domain	Structural analogy that licenses it here
Gini / Lorenz / Theil	economics of inequality	spread of recorded rape (and of charge rates) = spread of income across households
Funnel plot, control limits	clinical / institutional audit	is a force's charge rate an outlier? = is a hospital's mortality an outlier?
Binomial GLM, deviance accounting	statistics	how much does place vs offence drive a charge? = nested variance decomposition
Gradient boosting, calibration	machine learning	assign a charge probability to a case = supervised prediction with a checkable answer
Association-rule lift	retail data mining	failures co-occurring in a case = products co-purchased in a basket
Non-negative matrix factorisation	signal processing	many correlated failures = mixtures of a few latent sources
Absorbing Markov chain	stochastic processes	a case's decision pathway = a walk absorbed at an outcome
Mutual information	information theory	"does this risk foretell this failure?" = channel between two variables
k-core, betweenness, articulation	graph control theory	reform leverage = structural criticality of a node

Table 1: The borrowings and the structural analogy that licenses each. The claim in every row is identity of structure, not metaphorical resemblance; that is what makes the import legitimate rather than decorative.

resemblance, and that is what makes the borrowing legitimate.

Two boundaries complete the framework. First, the object of analysis is *record integrity* — how the institutional record was made and where it broke — and never the credibility of a person who reported harm; the latter is both out of scope and, as a computational target, actively harmful, and we exclude it by design. Second, where we draw arrows between variables we read them as *hypotheses*, not established causes: the data are observational, and a structural regularity is a candidate for a causal story to be tested, not a verdict.

5 RESEARCH QUESTIONS

1. **RQ1.** Do recorded sexual-offence counts measure the measurement regime rather than the underlying harm, and can a lens borrowed from the economics of inequality quantify how far that is true across jurisdictions?
2. **RQ2.** At national scale, across roughly two million rows of police-outcomes open data, what is the probability that a recorded sexual offence is charged or shed; how much of the variation in that probability is driven by the offence and how much by the police-force area; and can a calibrated model assign defensible outcome probabilities by location and offence?
3. **RQ3.** Does the institutional failure behind those outcomes have a recoverable struc-

ture — predictable bundles, latent archetypes, a decision-pathway grammar, structural leverage points — in the public review record; what do measurement, outcomes and mechanism together imply for reform of policing, law and government; and how far can a fully-local automated pipeline be trusted to do this reading?

6 METHODOLOGY

6.1 Research design

The study design is mixed and has three arms feeding one analysis. The first is a descriptive secondary analysis of official European recorded-crime statistics, used to test the measurement-regime account quantitatively. The second is a statistical-modelling arm that fits explanatory and predictive models of the system's response to a recorded offence on national police-outcomes open data. The third is a design-science extraction pipeline that converts public statutory reviews into a coded, anonymised dataset, on which the cross-domain battery of structural methods is then run. The three arms are joined in the discussion and in a single reform map. All code is released and the derived data regenerate offline from committed inputs.

6.2 Data sources and evidence base

Four data sources are used. (i) *Measurement*: Eurostat's `crim_off_cat` dissemination table, from which we take recorded rape (ICCS 03011) per 100,000 population for the most recent year with broad coverage, across 33 European states (Eurostat 2026). (ii) *Outcomes*: Home Office police-recorded-crime and outcomes open data (Home Office 2025; Home Office 2026) — roughly two million rows in total, comprising the outcome of every offence recorded in England and Wales in 2023/24 by force, offence and quarter against the modern twelve-group outcome framework (about 190,000 of them sexual offences), the older 2005/06–2013/14 sanction series, and Community-Safety-Partnership recorded volumes across 359 local areas for the decade to 2024. Summing the outcome groups within a force-offence-quarter cell recovers the recorded-offence denominator, so true outcome probabilities are computable. (iii) *Human-validated failure corpus*: eleven public reviews and inquiries spanning IICSA investigation reports, serious case and child safeguarding practice reviews, a domestic homicide review, and the Rotherham and Telford inquiries (Table 4 in the appendix), coded by hand and constituting the gold standard. (iv) *Scaled model-coded corpus*: 61 reviews from 22 publishers, fetched from public sources, from which only derived structural codes and provenance are retained — never the report text, which remains under its publishers' copyright.

6.3 Selection and sampling strategy

The eleven-case gold set is purposive: it was chosen to span jurisdiction-relevant review types, institutional settings (family, residential, religious, custodial, online-facilitated) and decades, so that recurring structure cannot be an artefact of a single setting. Inclusion required that a case be a public document concerning institutional handling of sexual offending; exclusion removed anything behind a paywall, anything subject to reporting restrictions, and any material identifying a complainant. The 61-case corpus applied the same public-source inclusion rule at scale, with exclusions logged (nine candidate URLs were unreachable and recorded as such). The asymmetry between a small validated set and a larger model-coded set is deliber-

ate and is treated as a feature of the design, not a defect to be hidden: it lets us measure the automated reading against the manual one.

6.4 Coding and controlled vocabulary

Each case is coded into four controlled fields — agencies involved, risk indicators present, an ordered list of institutional decisions, and record-integrity failure flags — drawn from fixed vocabularies of 21, 18, 15 and 15 terms respectively (Appendix A). Fixing the vocabulary is what makes cases commensurable and the downstream structural methods well-defined. The human coding was performed by the author against written definitions; the automated coding was performed by a fully local large language model (`liquid/lfm2.5-1.2b`) served through an OpenAI-compatible endpoint, constrained to the controlled vocabulary by a JSON-schema grammar and decomposed into one focused extraction per field, because a single nested schema collapsed to empty on a model this small.

6.5 Analysis methods I — modelling the outcome surface

The outcome arm fits models on aggregated force-by-offence-by-quarter cells using frequency weights, which is exact and fast. Each method is stated with the formula that travels with it.

What drives a charge, and by how much. For a recorded offence the binary event "charged" is modelled by logistic regression, $\operatorname{logit}\Pr(\text{charged}) = \alpha + \beta_{\text{force}} + \gamma_{\text{offence}} + \delta_{\text{quarter}}$, fitted as a binomial generalised linear model. Coefficients exponentiate to odds ratios — the multiplicative effect of an offence type or a force on the odds of a charge. To separate the contribution of *place* from that of *offence*, we fit nested models and compare their McFadden pseudo-$R^2 = 1 - \ell_{\text{model}}/\ell_{\text{null}}$: the increment when force is added to an offence-only model is the share of explained variation that is a postcode effect.

Which forces are genuine outliers. A force's charge rate $p_i = k_i/n_i$ is compared to the national rate $\bar{p}$ with a funnel plot (Spiegelhalter 2005): the exact binomial control limits at z standard errors, $\bar{p} \pm z\sqrt{\bar{p}(1-\bar{p})/n_i}$, narrow as the force volume n_i grows, so a small force far from $\bar{p}$ is noise but a large one is a signal. Forces outside the 99.8% limits ($z = 3.09$) are flagged as beyond chance. The between-force inequality in rates is summarised by a Theil index, the inequality counterpart of the Gini used for measurement.

Assigning a calibrated probability. For prediction we fit a gradient-boosting classifier (histogram-based) on the same features and read its drivers by permutation importance — the increase in log-loss when a feature is shuffled. The model is judged not by fit but by *calibration*: predicted charge probabilities are binned and compared to observed rates, and discrimination is summarised by the area under the ROC curve and the Brier score. A calibrated model is what licenses "assign a probability given location and offence".

6.6 Analysis methods II — reading the failure structure

The structural arm imports the cross-domain battery; each method is stated with the formula that travels with it, so the borrowing is exact.

Inequality of recording. For recorded-rate values $x_1, \dots, x_n$ across states, the Gini coefficient is

$$G \;=\; \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}}, \tag{1}$$

read off the Lorenz curve of cumulative recorded rape against the cumulative share of states. $G = 0$ is perfectly equal recording; $G \to 1$ is maximal concentration.

Failure bundles. For failures A, B with support $P(\cdot)$ estimated as corpus frequency, the lift of the rule $A \Rightarrow B$ is

$$\operatorname{lift}(A \Rightarrow B) \;=\; \frac{P(A \cap B)}{P(A)\,P(B)}, \tag{2}$$

with confidence $P(B \mid A) = P(A \cap B)/P(A)$. Lift above one means the pair co-occurs more often than independence predicts; we report rules with support ≥ 0.18 and lift ≥ 1.2.

Latent failure factors. Writing the case-by-failure indicator matrix as $V \in \{0,1\}^{m \times p}$, non-negative matrix factorisation seeks $W \geq 0$, $H \geq 0$ minimising $\|V - WH\|_F^2$ with k factors, so that each case is a non-negative mixture of k latent failure archetypes. We use $k = 3$.

The decision grammar. Decisions are states of a Markov chain estimated from consecutive pairs across the corpus. Splitting states into transient (T) and absorbing outcomes (A) gives the transition matrix in canonical form, from which the fundamental matrix $N = (I - Q)^{-1}$ yields the expected number of steps to absorption, $t = N\mathbf{1}$, and the absorption probabilities $B = NR$ — where a typical referral ends up, and how long it takes to get there.

Early-warning signal. For a binary risk X and a binary failure Y, the mutual information

$$I(X;Y) \;=\; \sum_{x,y} P(x,y)\,\log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3}$$

measures, in bits, how much knowing the risk reduces uncertainty about the failure; one bit is near-determination of a binary outcome.

Structural leverage. On the agency–decision co-occurrence network we compute the k-core (the maximal subgraph in which every node has degree at least k) and the articulation points (nodes whose removal disconnects the graph). We then define a *structural leverage index* that combines a node's brokerage — its betweenness centrality C_B — with its prevalence f,

$$C_B(v) \;=\; \sum_{s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \qquad\qquad L(v) \;=\; C_B(v) \cdot f(v), \tag{4}$$

where $\sigma_{st}(v)$ counts shortest paths through v and $f(v)$ is the fraction of cases in which the node appears, so that L ranks where a reform would touch both the most cases and the most structurally pivotal point.

6.7 Analysis methods III — simulating repair

The reform arm asks counterfactual questions, and answers them with simulation rather than assertion; each rests on stated, deliberately conservative assumptions and carries its uncertainty. (i) An *attrition cascade*: a recorded rape flows recorded $\rightarrow$ charged $\rightarrow$ convicted, with the charge probability drawn from the data as a Beta posterior and the charge-to-conviction probability from a stated prior (mean 0.6); twenty thousand cohorts are simulated under the baseline and under three reform scenarios — levelling charging up to the top-quartile force, supporting complainants so a fraction (prior mean 0.18) of withdrawn cases reach a charge, and both — and the convictions per thousand recorded offences are reported as medians with 90% intervals. (ii) A *force-levelling counterfactual*: for a target percentile q of the force charge-rate distribution, every below-target force is raised to that rate and the additional annual charges are summed. (iii) A *decision-grammar reform*: cohorts are simulated through the review-derived absorbing Markov chain under the empirical transition matrix and under an early-recognition reweighting that multiplies transitions toward protective and forward states by $(1+\delta)$ and toward premature closure by $(1-\delta)$, and the absorbing outcomes are compared. These are "what-if" arithmetic, not forecasts, and are read as such.

6.8 Validity, reliability and trustworthiness

Reliability of the automated reading is quantified against the human gold standard with Cohen's κ, percentage agreement and mean Jaccard overlap per field (Table 2). Robustness of the structural findings is addressed by running the failure-level analytics on the validated set and the network and pathway analytics on the larger corpus, and by reporting where the two diverge. The methodological limitations are stated in full in Section 12; the most important is that the failure-level analytics rest on a small validated corpus and are read as exploratory structure rather than confirmed prevalence.

7 FINDINGS

7.1 Act I — the numbers mislead, predictably

Recorded rape per 100,000 population varies across European states by more than an order of magnitude, with the highest-recording states an order above the median (Figure 1). A naive reading treats this as a danger ranking. The inequality lens shows why that reading is untenable. The Lorenz curve of recorded rape (Figure 2) bows far from the line of equal recording, with a Gini coefficient of $G = 0.52$. To put that number in human terms: recorded rape is spread across European states about as unequally as household income is spread across households in Brazil. Income is genuinely that unequal; recorded rape is not, because the quantity being concentrated is not danger but the propensity of a regime to define broadly, count by victim, and record diligently. Sweden, whose consent-based definition and per-incident counting are designed to capture more, sits at the top by construction, not by catastrophe (Office for National Statistics 2021). The same mechanism explains the discontinuities in national time series: when Germany broadened its offence around 2016 its recorded figure stepped up by roughly a quarter without any claim that the country had suddenly become more dangerous (Figure 3). The measurement-regime account is therefore not a caveat to be noted and set aside; it is the single largest signal in the cross-national data, and it answers RQ1 in the affirmative.

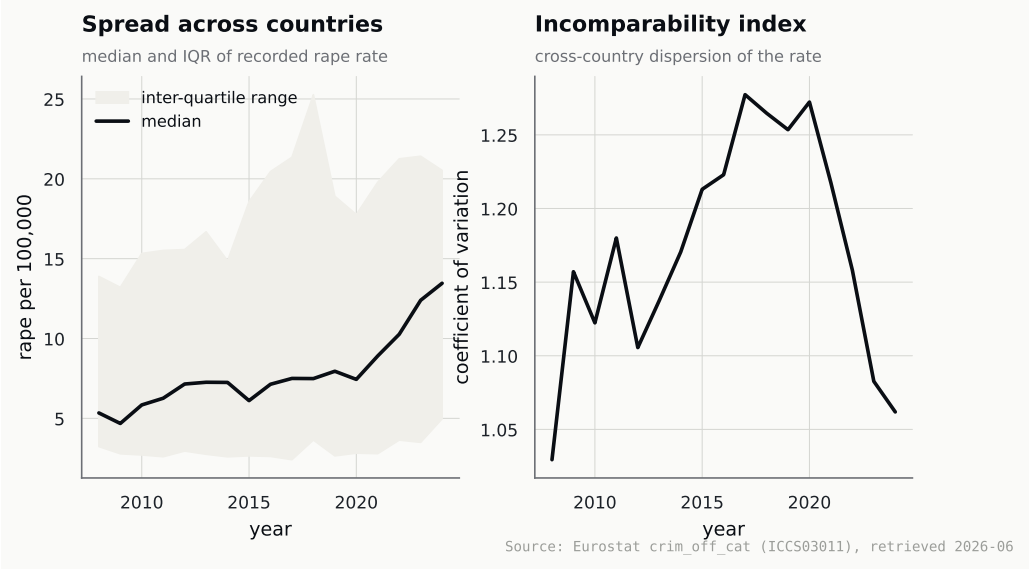


Figure 1: Recorded rape per 100,000 across European states (latest year with broad coverage). The spread exceeds an order of magnitude. Read as a ranking of measurement regimes — definition breadth, counting rule, recording propensity — not of danger.

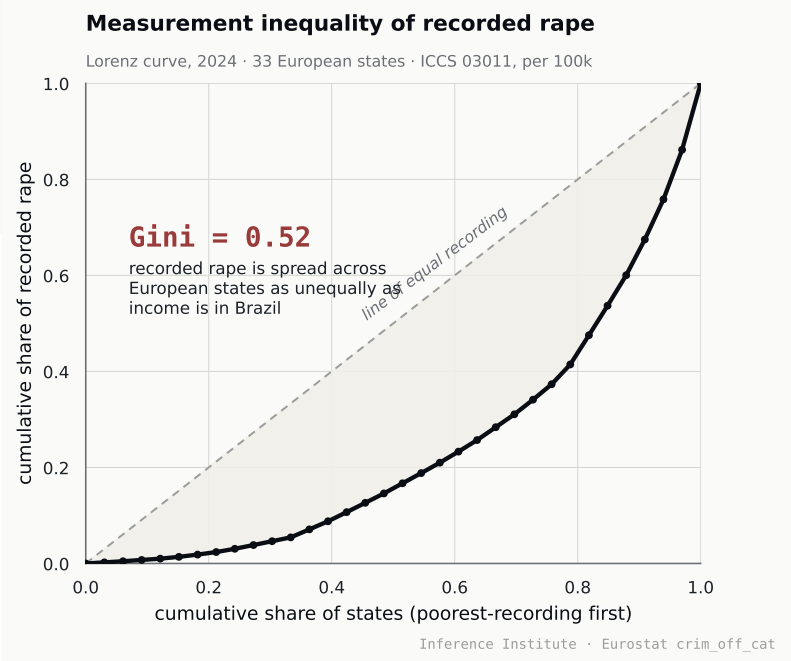


Figure 2: The same data as a Lorenz curve. The area between the curve and the diagonal gives a Gini coefficient of 0.52: recorded rape is concentrated across states about as unequally as household income in a highly unequal economy. Because the concentrated quantity is recording propensity, the coefficient measures the incomparability of the counts, not a gradient of harm.

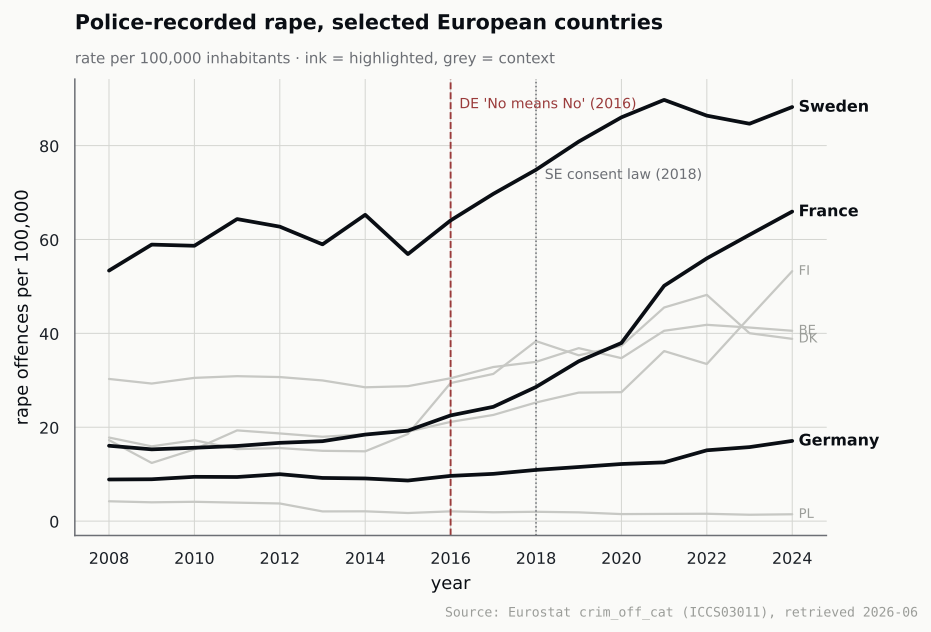


Figure 3: National time series for selected states. Step changes coincide with legal-definition and recording reforms rather than with plausible changes in underlying incidence; the German step around 2016 is the clearest example.

7.2 Act II — the outcome surface

The second movement follows a recorded offence to what the system does with it. Where the review record holds a few dozen catastrophes in depth, the Home Office outcomes open data holds the shallow truth about all of them: the outcome assigned to every one of the roughly 190,000 sexual offences recorded in England and Wales in 2023/24, by force, offence and quarter. Modelled, it draws an *outcome surface* — the probability that a recorded offence meets each fate — and that surface is steep, uneven, and predictable. The title of this paper is the end of that surface (Figure 4): of a thousand recorded rapes, about thirty-five are charged and, on a stated charge-to-conviction rate, near twenty-one convicted — roughly three in a hundred, and that is before counting the offences never reported at all.

The modal response to a recorded rape is not a charge. The headline is stark (Figure 5). Of recorded rapes, only 3.5% result in a charge or summons; 49.7% are closed under the outcome "evidential difficulties — victim does not support further action," and a further 15.3% are closed for evidential difficulties despite the victim supporting action. For other sexual offences the charge rate is 6.9% and the victim-withdrawal share 30.8%. Read plainly: the single most likely recorded outcome of a reported rape is not a prosecution, a caution, or even a finding that no suspect could be identified — it is that the case ends because the complainant, somewhere in a long process, stops supporting it. This is the system's own recorded account of itself, and it locates the attrition not at the courtroom door but far upstream, in whatever happens to complainants between report and charging decision.

Where you report changes the odds. That national average conceals a wide geography. The force charge rate for sexual offences ranges from 2.8% to 11.0% — a near-fourfold spread — and this is not the noise of small forces. A funnel plot (Figure 6), the tool clinical audit uses to separate signal from sampling variation, places *eighteen of forty-four* forces outside the 99.8% control limits around the national rate: nine charge significantly more often than chance would allow, nine significantly less. The between-force inequality has a Theil index of 0.04, the outcome-side echo of the measurement inequality of Act I. The fitted model turns this into a defensible per-force probability with its uncertainty (Figure 7): the predicted chance that a recorded rape is charged runs from about one in fourteen in the highest force to about one in fifty in the lowest, with confidence intervals that for many forces do not overlap. Two complainants reporting the same offence in different force areas do not face the same system.

The same geography governs where complainants are lost. The victim-withdrawal closure rate is even more variable than the charge rate — *twenty-seven* of forty-four forces sit outside its funnel's control limits (Figure 8), from around one in six to more than half. Because withdrawal is the largest single outcome, this second lottery is where the most ground is to be gained or lost.

Offence type leads, but place is the second driver. What drives the charge decision? A deviance decomposition is candid about scale: offence type alone accounts for a McFadden pseudo-R^2 of 0.043, the police force alone for 0.009, and the two together for 0.051 — so most of the variation between individual offences is irreducible at this resolution, as one would expect when outcomes turn on case facts the data do not contain. But the systematic part has a clear order. A gradient-boosting model agrees with the GLM (Figure 9a): offence type is

Figure 4: The national attrition funnel for rape, per 1,000 recorded offences in 2023/24. Each step sheds the great majority of what reached it; the upstream estimate (faded) assumes roughly one in six offences is reported. "Three in a hundred" is the charge step; the conviction step is read off a stated charge-to-conviction prior.

the strongest driver of the charge probability, and the police force — pure location — is the second, ahead of the quarter and well ahead of the rape-versus-other distinction, which the offence code already carries. Adjusting for offence and quarter, recorded rape has about half the odds of a charge of other sexual offences (odds ratio 0.51). The model is not merely suggestive: it is calibrated (Figure 9b), its predicted charge probabilities tracking observed rates across the full range, with an ROC area of 0.73 and a Brier score of 0.05. That calibration is what licenses the practical claim — that one can assign a defensible charge probability to a recorded offence given its type and its force — without overclaiming a causal account.

A large and rising base. This steep, uneven surface is applied to a growing volume. Across the Community-Safety-Partnership open data for the decade to 2024, recorded sexual offences roughly doubled between 2015/16 and 2021/22 before plateauing (Figure 10), driven by both rising disclosure and broadened recording. The low charge rate is therefore not a small system clearing a small backlog; it is the standing response to a large and rising number of reports. Together these results answer RQ2: the outcome surface is steep (few charges), shed mainly through victim withdrawal, geographically unequal beyond chance, driven first by offence and second by place, and modellable to a calibrated probability. It is, as the next act argues, the aggregate shadow of a mechanism.

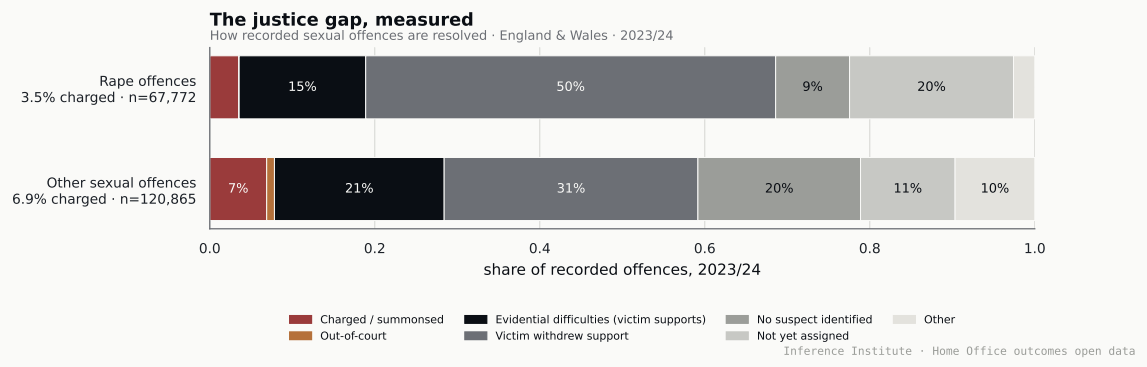


Figure 5: How recorded sexual offences are resolved in 2023/24. The charge slice is small for both groups and smallest for rape (3.5%); victim withdrawal is the largest single outcome for rape (49.7%). The system's modal recorded response to a reported rape is to close it for want of a supported complaint.

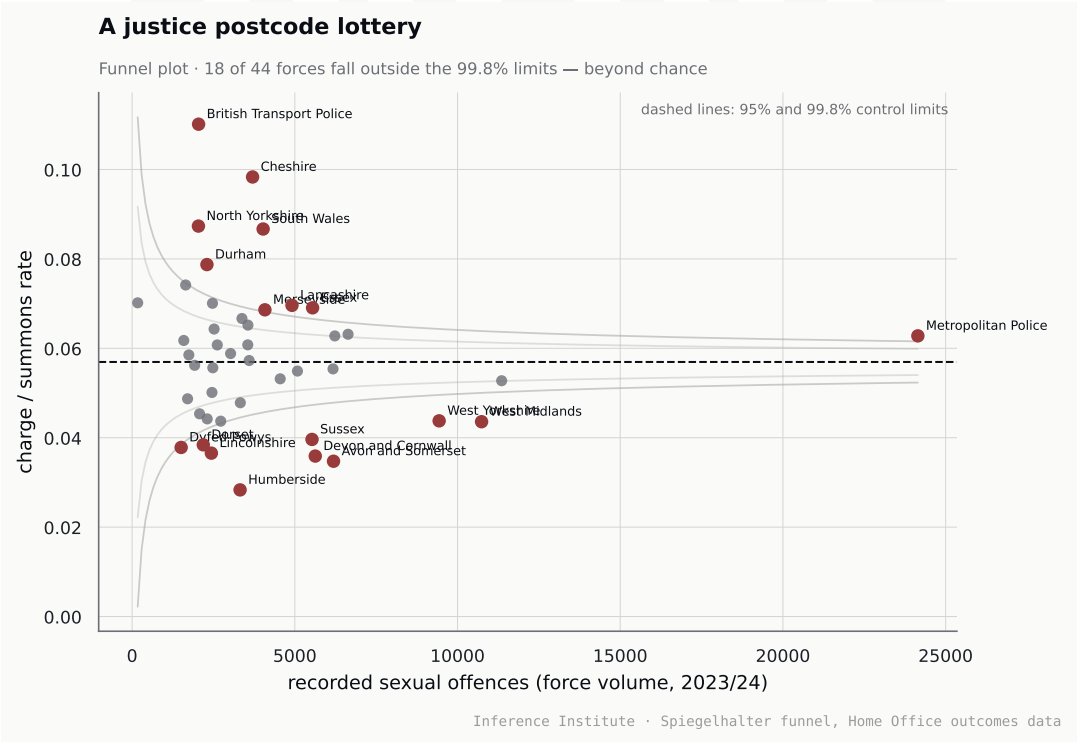


Figure 6: Funnel plot of force charge rates against force volume, with 95% and 99.8% binomial control limits around the national rate (5.7%). Eighteen of forty-four forces fall outside the 99.8% limits — genuine outliers, not noise. The justice an offence meets depends measurably on where it is reported.

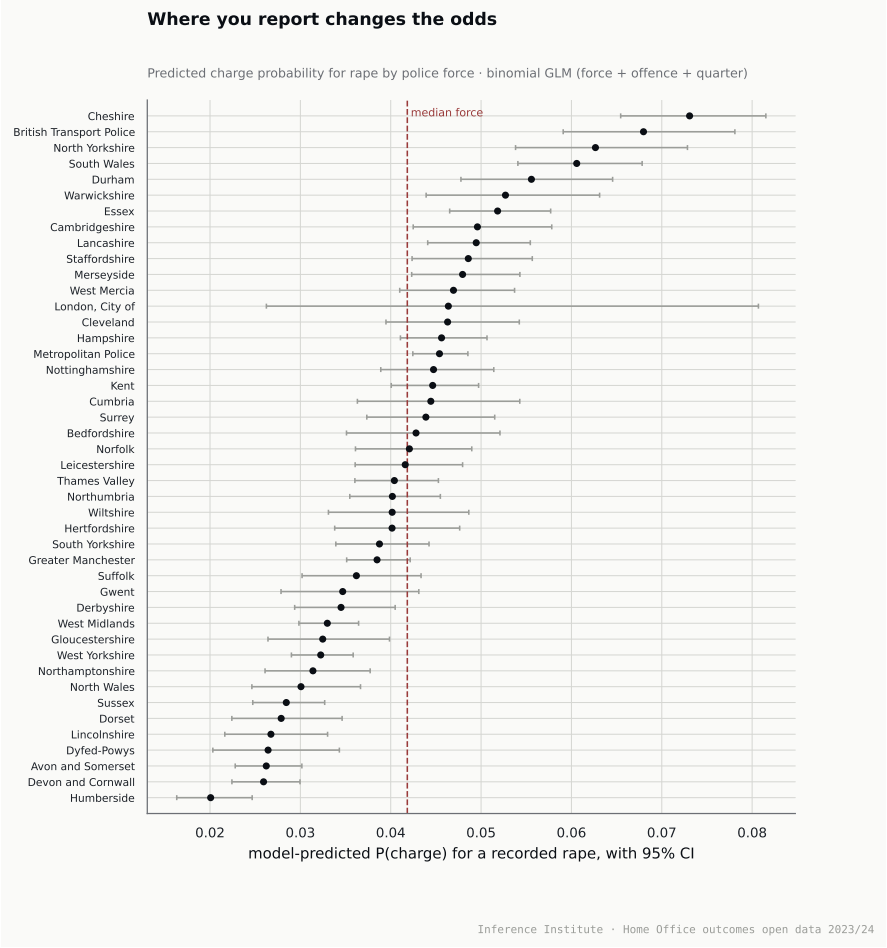


Figure 7: Model-predicted probability that a recorded rape is charged, by force, with 95% confidence intervals (binomial GLM, force $+$ offence $+$ quarter). The dashed line is the median force; the spread and the many non-overlapping intervals are the postcode lottery made precise.

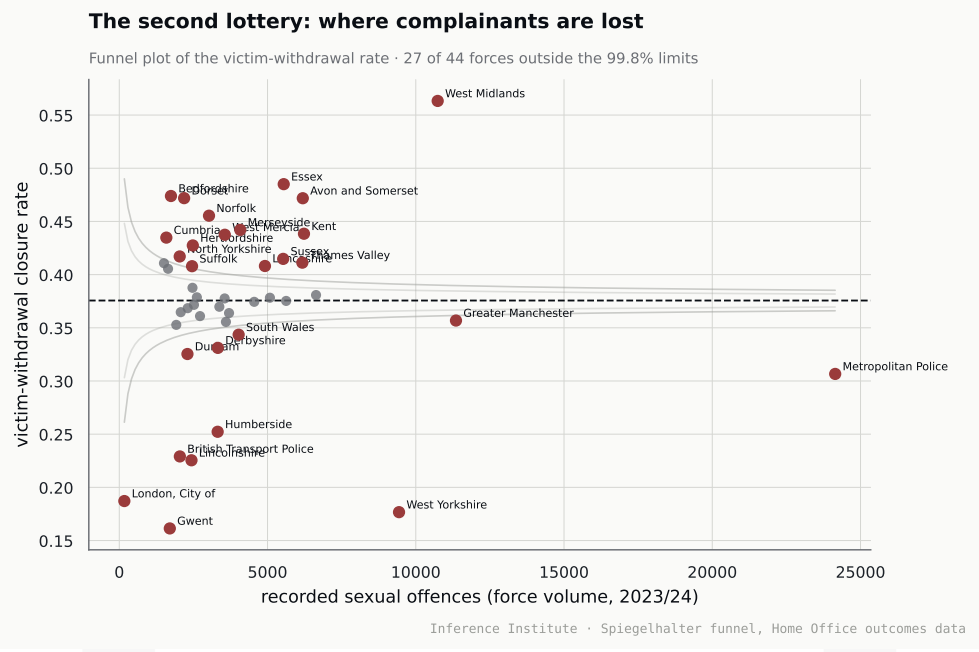
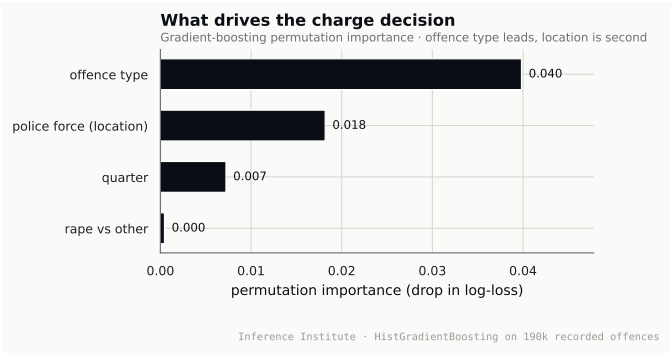
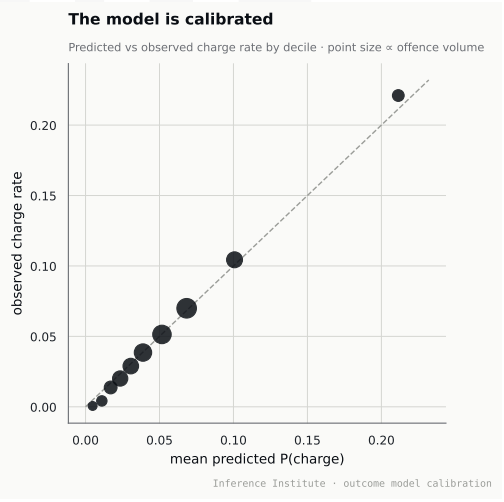


Figure 8: Funnel plot of the victim-withdrawal closure rate by force. It varies even more than the charge rate, with 27 of 44 forces beyond the 99.8% control limits. "Victim does not support action" is the police-recorded outcome category, read here as a property of the *process* that loses complainants, never of the complainant.



(a) Drivers (permutation importance)



(b) Calibration

Figure 9: (a) Offence type is the strongest driver of whether a recorded offence is charged; the police force — location — is second. (b) The model is well calibrated: predicted and observed charge rates agree across deciles, so the per-force probabilities can be trusted as probabilities.

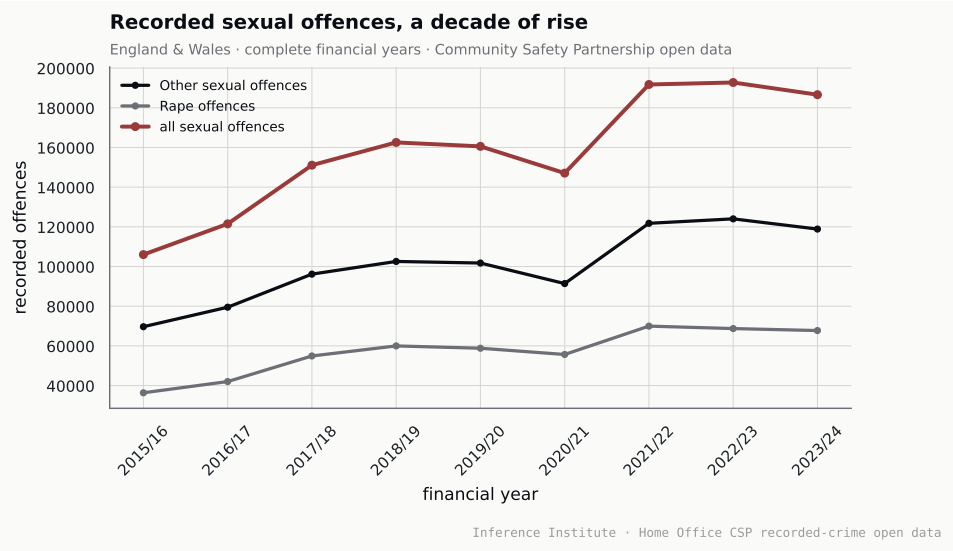


Figure 10: Recorded sexual offences in England and Wales over the decade to 2024 (complete financial years, Community-Safety-Partnership open data). The base to which the low charge rate applies roughly doubled and then plateaued.

7.3 Act III — failure has a grammar

The third movement asks *why* the outcome surface looks as it does, and turns from the aggregate to the mechanism written in the review record. The question is whether that record has a structure a coded reading can recover. It does, in five complementary ways.

Failure arrives in predictable bundles. Treating each case as a basket of record-integrity failures and mining association rules surfaces strong, interpretable regularities (Figure 11). Where recruitment and vetting failed, prior reports or reviews had already been ignored, every time it occurred (lift 3.67, confidence 1.00): the institution had been told. Where cases were handled in isolation, information was not shared (lift 2.20, confidence 1.00). And wherever the record showed reputation being placed above protection, an allegation had gone unacted-upon (lift 1.83, confidence 1.00). These are not platitudes; they are conditional structure — the presence of the antecedent makes the consequent far more likely than chance — and each is a candidate early-warning rule: when an inspector sees the antecedent, the consequent is where to look.

Many failures, three hidden causes. Non-negative matrix factorisation decomposes the fifteen failure types onto three latent factors that account for the observed co-occurrence with low reconstruction error (Figure 12). The first and heaviest factor loads on allegations not acted upon, reputation placed over protection, inadequate oversight, procedures ignored and leadership failure — an *institutional self-protection* archetype. The second loads on risk underestimated, victim-blaming language, failure to escalate, information not shared and crime not recorded — a *recognition-and-recording* archetype. The third loads on leadership failure, ignored reports, poor vetting and inadequate oversight — an *oversight-and-vetting collapse*. Cases are mixtures: the religious, residential and custodial settings load most heavily on self-protection, while the child-sexual-exploitation cases (Rotherham, Oxfordshire, Telford) carry more of the recognition-and-recording factor. That the dominant latent cause across the corpus is self-protection, not incompetence, is the single most consequential structural finding for reform.

The grammar of a case. Modelled as an absorbing Markov chain, the decision pathway has a definite shape (Figure 13). Estimated over the 61-case corpus, a typical referral is absorbed into a protective outcome (a prosecution or conviction) with probability 0.39, into a retrospective external review with the same probability 0.39, and into no further action with probability 0.22, after an expected ten decision steps. The symmetry is the point. In the documented record of how these cases proceed, a reviewed case is as likely to end with the system being *told to look again* as with it acting protectively of its own motion. The external review is not an epilogue to the process; it is one of its two dominant terminal states. Recognition, in this grammar, arrives late and often only when forced — a reading reinforced by the centrality analysis below, in which the external review is the structurally most central decision of all (Figure 14).

A dense core with no single point of failure. The agency–decision co-occurrence network over the corpus is dense: 32 nodes joined by 241 links, with a 13-core — twenty nodes every one of which co-occurs with at least thirteen others (Figure 15). Critically, it has *zero* articula-

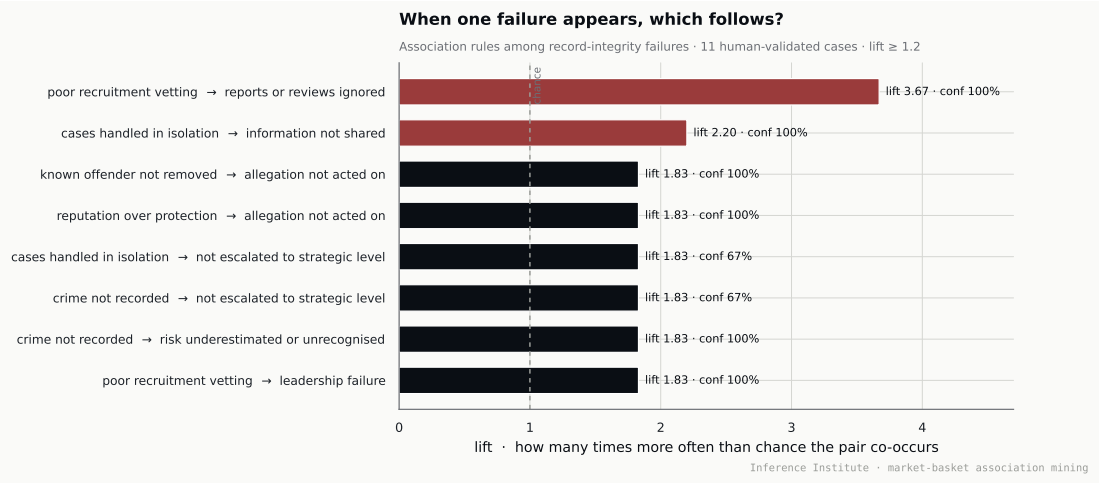


Figure 11: Association rules among record-integrity failures (11 validated cases). Bars show lift — how many times more often than chance a pair co-occurs — with confidence annotated. The strongest rules are read as candidate early-warning relationships: the antecedent is a visible cue, the consequent is the hidden failure it predicts.

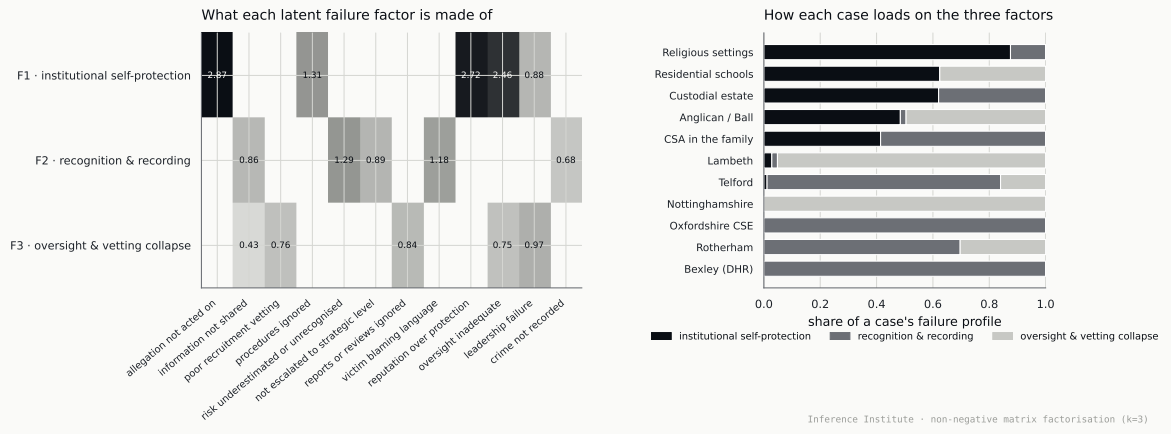


Figure 12: Latent failure factors (non-negative matrix factorisation, $k = 3$). Left: what each factor is made of, as weights on individual failures. Right: how each named case loads on the three factors. Failure is not fifteen independent problems but three recurring archetypes, dominated by institutional self-protection.

tion points: there is no agency whose removal would disconnect the response, because every relevant agency is already connected to every other. The brokers with the highest betweenness are the procedural junctions — the investigation and the referral (each $C_B = 0.079$), then the police (0.046) and local-authority leadership (0.043). This is the structural refutation of the "work in silos" diagnosis. The institutions were not disconnected; the infrastructure of connection was already complete and redundant. What failed was not the link but the *signal* the link was supposed to carry. Reform that adds connection therefore targets a problem the data say does not exist; reform must instead change what the existing, dense connections transmit and who is accountable at the broker junctions.

Which risks foretell which failures. If connection is not the problem, signal is, and information theory locates exactly where the signal is strongest (Figure 16). The presence of an abuse-of-position risk, or of an organisational-culture risk, is almost perfectly informative about an oversight-inadequate failure: mutual information of 0.99 bits, meaning the risk all but de-

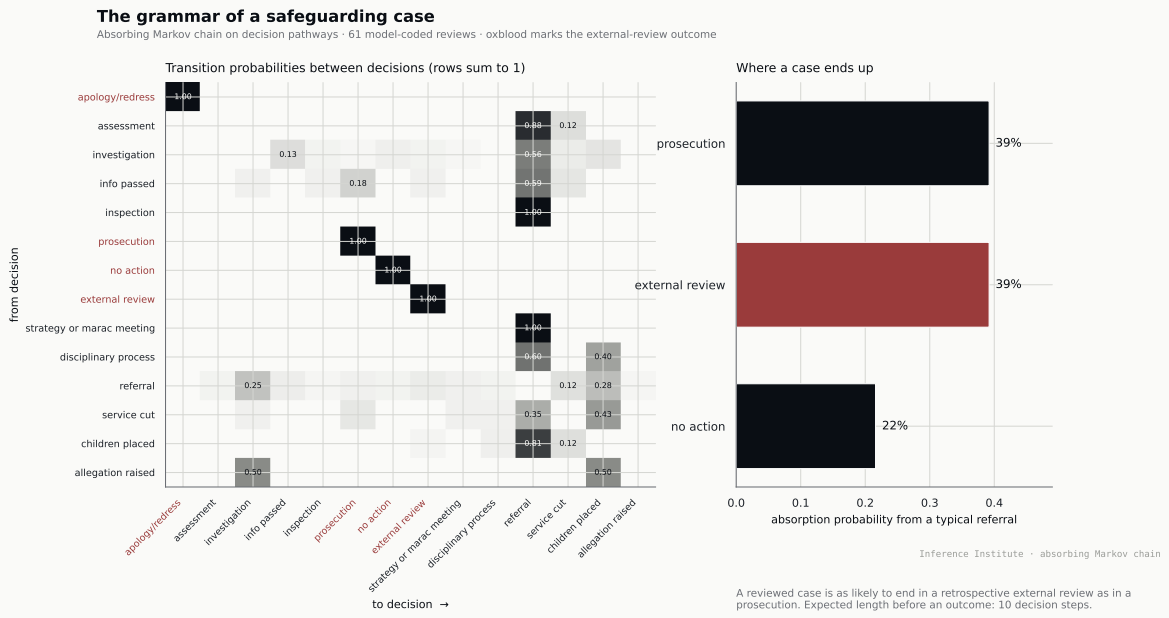


Figure 13: The decision grammar as an absorbing Markov chain (61 model-coded reviews). Left: transition probabilities between decisions. Right: where a typical referral is absorbed. A reviewed case ends in a retrospective external review as often as in a prosecution, after about ten decision steps — evidence that the system most reliably recognises harm only once an inquiry compels it.

termines the failure. A residential setting strongly foretells a recruitment-and-vetting failure (0.68 bits); a going-missing risk foretells an unacted-upon allegation (0.62 bits). These are the points at which the existing dense network already *contains* the predictive signal — the risk is recorded — but does not act on it. The combination of the two findings is the heart of the paper: the system has the connections and it has the signal; what it lacks is transmission and recognition.

These five readings answer the structural half of RQ3. Failure is not a formless catalogue of mishaps. It comes in predictable bundles, it is generated by three latent archetypes, it follows a decision grammar with a definite absorbing structure, it exposes its leverage at identifiable broker junctions, and it is foretold, sometimes almost deterministically, by risks the record already holds. The co-occurrence of failures separates cleanly into syndromes (Figure 17a), and the semantic content of the reviews is only weakly aligned with their structural form (Figure 17b, correlation $r \approx 0.25$), which is why a structural reading recovers patterns a textual one misses.

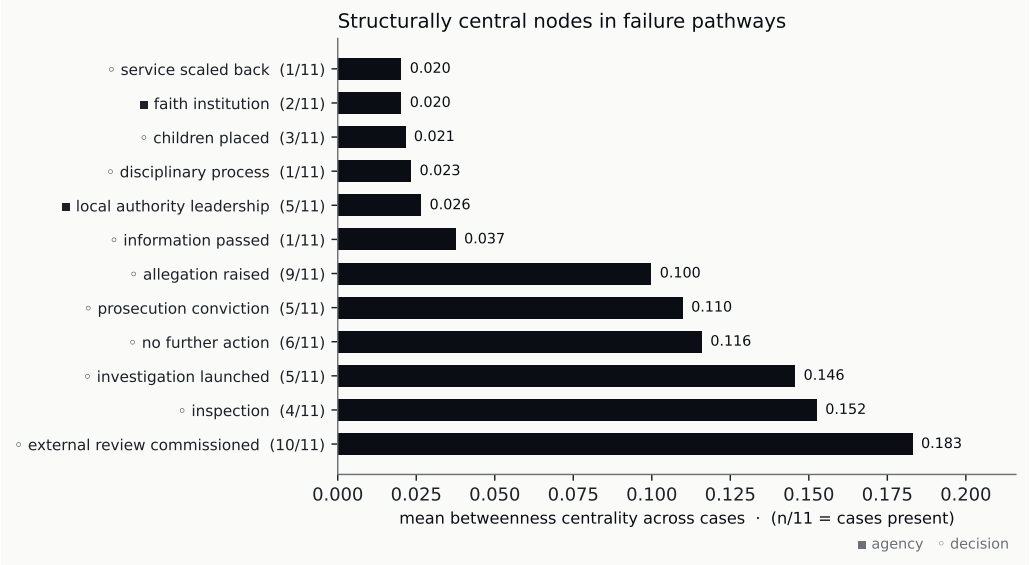


Figure 14: Structurally central nodes in the failure pathways (mean betweenness across cases). The external review is the most central decision on this and all other centrality measures — the hub through which the system's response is organised, after the fact.

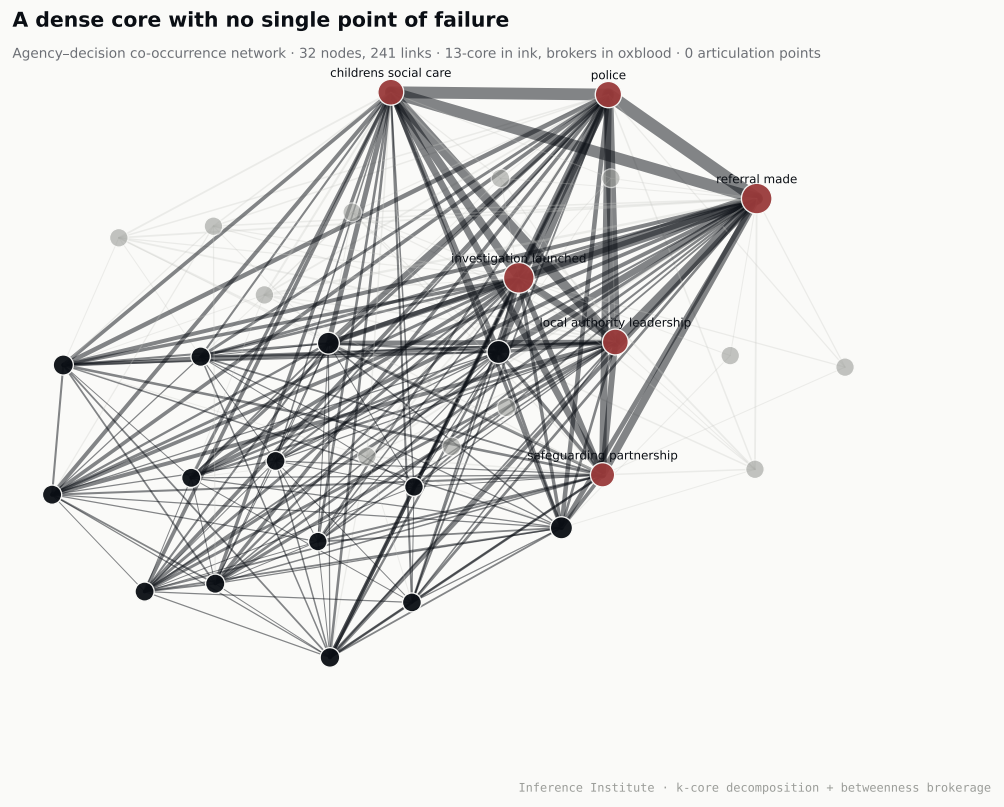


Figure 15: The agency–decision co-occurrence network (k-core in ink, brokers in oxblood). The network is a dense 13-core with no articulation points: there is no missing connection to add. The leverage lies at the high-betweenness procedural junctions — referral and investigation — not in new information-sharing infrastructure.

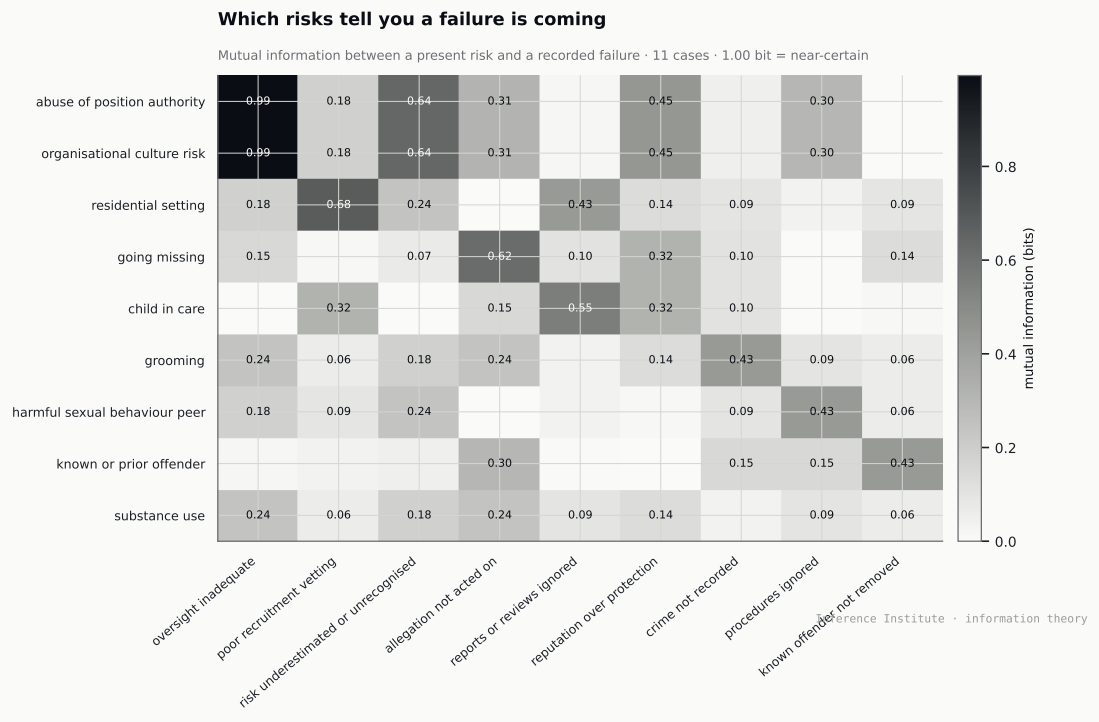
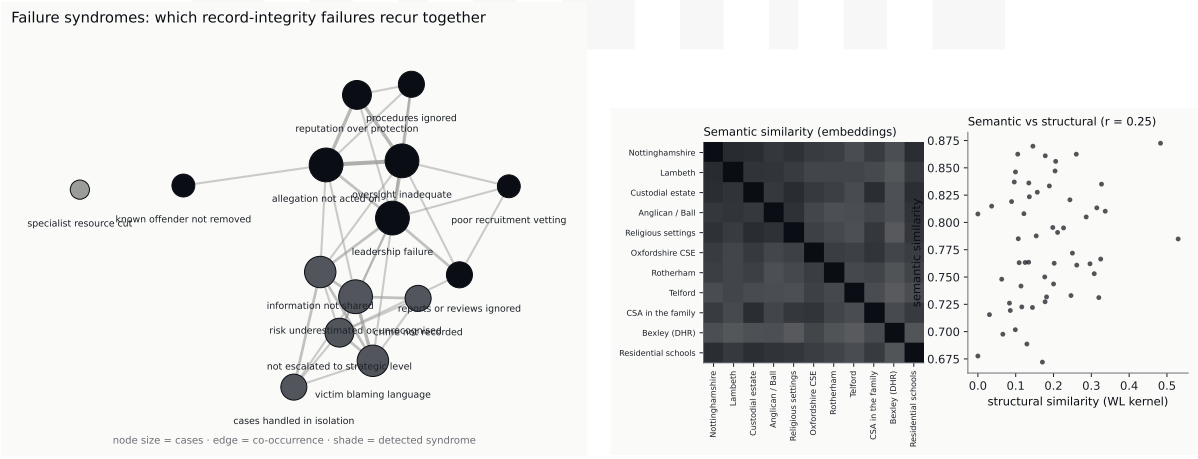


Figure 16: Mutual information between a present risk and a recorded failure (11 validated cases). An abuse-of-position or toxic-culture risk is near-deterministic of an oversight failure (≈ 1 bit). The predictive signal is present in the record; the failure is that it is not acted upon.



(a) Failure syndromes **(b)** Semantic vs. structural similarity

Figure 17: (a) Record-integrity failures cluster into distinct syndromes — groups that travel together. (b) Semantic similarity (from local embeddings of the review text) is only weakly correlated with structural similarity ($r \approx 0.25$), so how a review *reads* is a poor guide to how a case *failed*.

7.4 The honest machine

RQ3 also asks how far the automated reading can be trusted, and the answer is specific rather than rhetorical (Table 2). Against the human gold standard, the fully-local 1.2-billion-parameter model recovers *which agencies* were involved tolerably well (Cohen's $\kappa = 0.60$), because agencies are named explicitly in the text. It recovers the interpretive coding poorly — risk indicators ($\kappa = 0.23$), integrity flags ($\kappa = 0.19$) and especially the ordered decision types ($\kappa = 0.07$) — because those require judgement the small model does not reliably exercise. At scale the failure is diagnostic: the model tags "cases handled in isolation" in almost every review, an artefact of its prior rather than a property of the corpus. The consequence for this paper is a discipline we adopt openly: the agency and pathway structure, where the model is sound, is read on the 61-case corpus; the interpretive failure structure, where it is not, is read only on the eleven human-validated cases. The machine is useful exactly to the extent that we can measure, and bound, its unreliability.

Category	Cohen's κ	% agreement	Mean Jaccard
agencies	0.60	0.85	0.50
risk_indicators	0.23	0.73	0.26
integrity_flags	0.19	0.66	0.27
decision_types	0.07	0.63	0.17

Table 2: Reliability of the fully-local model against human coding, by field. Structural facts that are written in the text (agencies) are recovered; interpretive judgement (decisions, flags) is not. This is why the interpretive analytics in this paper rest on the validated corpus, not the scaled one.

8 SIMULATING REPAIR

The first three movements describe how the system works now and diagnose why it fails. Reform asks a different, counterfactual question — *what would change if?* — and that question is answered badly by assertion and well by simulation. We run three experiments on the same data and models, each with stated, deliberately conservative assumptions and each reported with its uncertainty. They are what-if arithmetic, not forecasts.

8.1 The cascade, and the single biggest lever

Treating a recorded rape as a flow — recorded $\rightarrow$ charged $\rightarrow$ convicted — and drawing the charge probability from the data and the charge-to-conviction probability from a stated prior, twenty thousand simulated cohorts put the baseline at about **21 convictions per 1,000 recorded rapes** (Figure 18). The interesting part is the comparison of levers. Levelling every force up to the top-quartile charge rate raises that to about 40 ($\times 1.9$). *Supporting complainants* so that even a modest fraction of withdrawn cases are carried to a charge raises it to about 74 ($\times 3.5$) — nearly twice the effect — and doing both reaches about 93 ($\times 4.4$). The reason is structural, not arithmetic luck: victim withdrawal is the largest single outcome, so the lever that acts on it has the most mass to move. The simulation thus turns the descriptive finding of Section 7.2 into a priority ordering — complainant support first — with an honest 90% interval around every bar and its assumptions (a 0.6 conviction-of-charge rate, 0.18 of withdrawals recoverable) printed on the figure.

8.2 Closing the postcode lottery

The geography of charging is an injustice, but it is also an opportunity, because the gap between the laggard forces and the rest is charges that are currently lost to location alone. A deterministic counterfactual makes the size plain (Figure 19): if every below-target force charged at the median force's rate, England and Wales would record roughly **1,200 additional sexual-offence charges a year**; at the 75th percentile, about $2{,}300$; at the 90th, about $4{,}000$ — an uplift of 11 to 38% on the current total, with no change in the law and no new offence, simply by closing the spread the funnel plot identifies as beyond chance.

8.3 Moving recognition upstream

The mechanism analysis said the system recognises harm late and only when forced; the leverage index said to act at the referral and investigation junctions. The third simulation tests what that would do to the decision grammar itself. Re-weighting the review-derived transition matrix toward protective and forward states and away from premature closure —

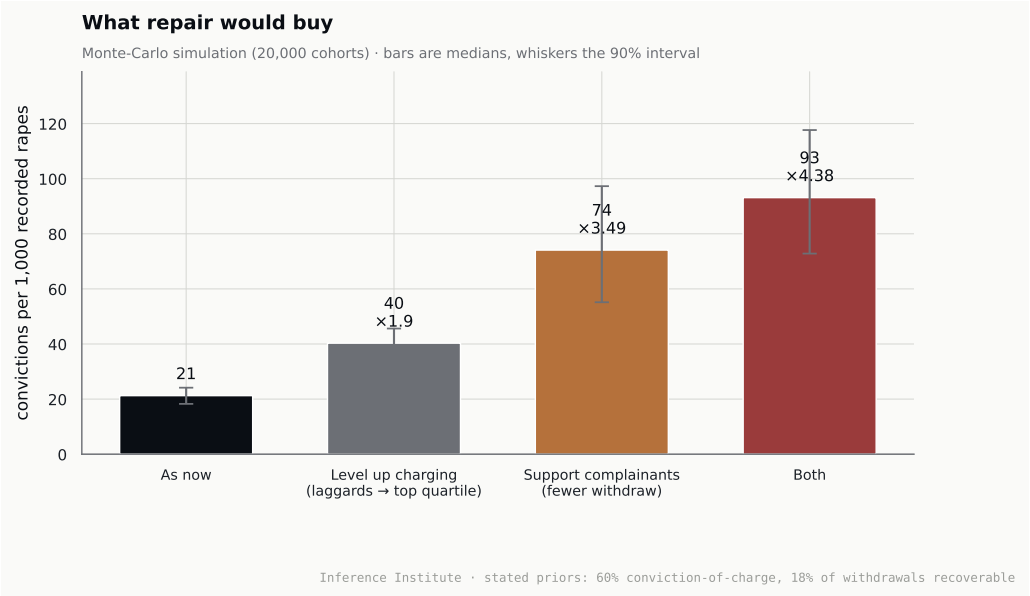


Figure 18: Convictions per 1,000 recorded rapes under the status quo and three reform scenarios, from 20,000 Monte-Carlo cohorts (medians; whiskers are 90% intervals). Supporting complainants is the single largest lever because victim withdrawal is the dominant outcome. Priors are stated and deliberately conservative; this is what-if arithmetic, not a forecast.

a transparent, single-parameter intervention — and simulating fresh cohorts shifts the share of cases absorbed into prosecution from **39% to 50%**, cuts the share ending in no further action from 21% to 14%, and shortens the average pathway from ten decision steps to nine and a half (Figure 20). The reform that the structural analysis pointed to is, in simulation, exactly the one that moves cases from the retrospective external review — the system's late, forced recognition — into a timely protective outcome.

Together the three experiments convert the reform map from a list of good ideas into a set of quantified, falsifiable predictions: each reform now carries an expected effect, a direction, and an interval, and each is testable against the natural experiments — above all the staggered rollout of Operation Soteria — named in Section 13. That is the difference between asserting that a system could be better and showing, on its own data, by how much.

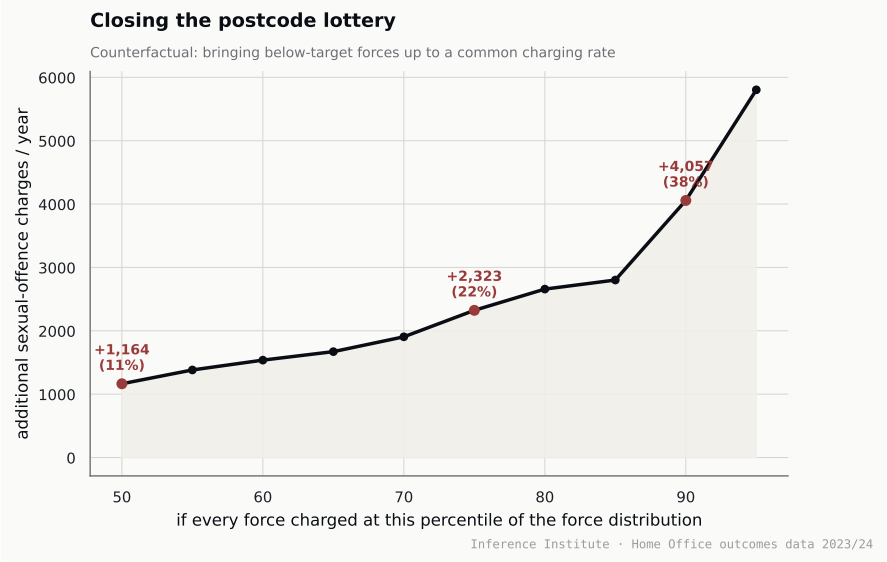


Figure 19: Additional annual sexual-offence charges if every below-target force were brought up to a common charging rate, as a function of the target percentile of the current force distribution. The marked points are the median, 75th and 90th percentiles.

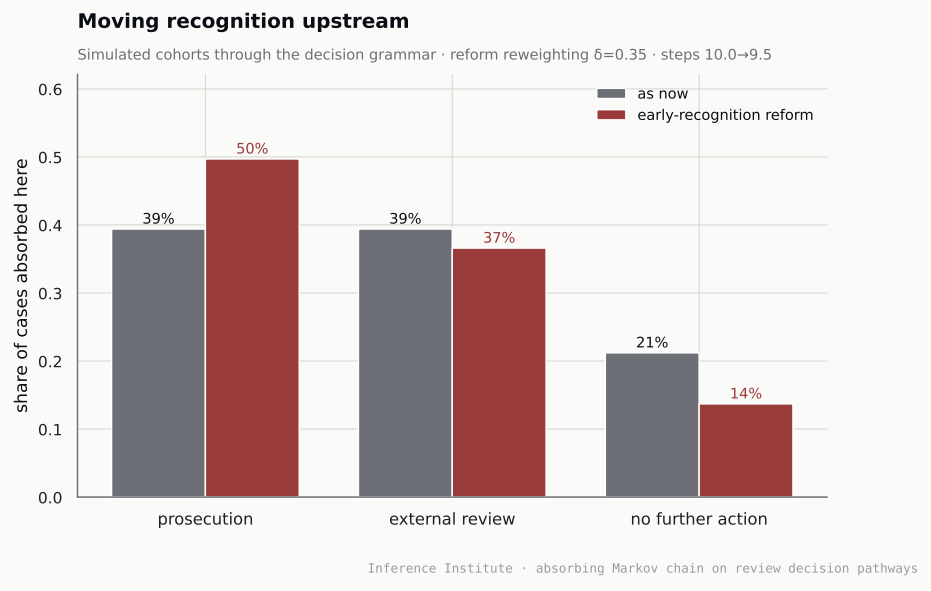


Figure 20: Where simulated cases are absorbed in the decision grammar, now versus under an early-recognition reweighting ($\delta = 0.35$). Reform moves cases out of no-further-action and the retrospective external review and into prosecution, and shortens the pathway.

9 DISCUSSION

Read together, the findings tell one story, and it is not the story the review record tells about itself. The standing narrative is one of fragmentation: agencies that did not talk to each other, information that fell between services, silos to be dismantled. The structural reading contradicts it at the decisive point. The agencies were densely connected — a 13-core with no single point of failure — and the predictive signal was present, sometimes to the point of near-determination. The network had neither a missing link nor a missing measurement. What it lacked was *transmission and recognition*: a dense, redundant set of connections that did not carry the signal they contained, and a process whose dominant moment of recognition was a retrospective external review, arriving after about ten decision steps, as often as a protective outcome arrived at all.

This reframes the problem from connectivity to signal, and it is a reframing with teeth, because the two diagnoses imply opposite reforms. The connectivity diagnosis funds infrastructure — another database, another hub, another duty to cooperate — at junctions that are already saturated. The signal diagnosis funds something harder and cheaper: accountability for acting on what is already known, concentrated at the broker junctions the leverage index identifies, and a legal and cultural assault on the self-protection factor that the latent-cause analysis shows to be the dominant generator of failure. The mutual-information result tells a reformer where the signal is loudest and most ignored; the association rules tell an inspector what to look for when a cue appears; the Markov grammar tells a regulator that the system's recognition is late by design and must be moved upstream; and the leverage index tells a minister which junction to make accountable first.

The outcome surface is the same story read at national scale, and it makes the mechanism quantitative. A system that recognises harm late and protects itself first will, in aggregate, shed most of what it records — and it does: 3.5% of recorded rapes charged, half closed because the complainant withdraws. Victim withdrawal is not a fact about complainants; it is the recorded signature of a process that loses their support, and its dominance is what the self-protection and late-recognition mechanism looks like spread over 190,000 cases. The postcode lottery sharpens the point. If the outcome were governed only by case facts and the law, it would not depend on the force; that eighteen of forty-four forces sit outside the funnel's control limits, with location the second-strongest driver after offence type, says the mechanism is not a uniform property of "the system" but something forces do differently — which is precisely why it can be reformed, and why a calibrated per-force probability is a management instrument, not merely a description. The review record tells us *how* a case fails; the outcome surface tells us *how often* and *where*; the two are one argument at two scales.

The measurement half of the paper is not a separate argument but the same one at the level of numbers. A system that recognises harm only when forced will also count it according to whatever regime it happens to run, and will then compare those counts as though they measured danger. The Gini result is the macro-scale shadow of the micro-scale finding: in both, an artefact of process is being read as a fact about the world. Honest reform has to fix the reading in both places — harmonise the measurement before ranking states, and change the signal transmission before adding more connection.

10 CONTRIBUTION: A STRUCTURAL MAP FOR REFORM

The paper's constructive contribution is to convert each structural finding into an ownable reform, via the leverage index of Equation 4 and the map in Table 3. This is the "map" the title promises: not a wish-list, but a correspondence in which every reform is anchored to a measured structure and assigned to the level of the system that could enact it.

Contribution type: Method, taxonomy, model and empirical finding

Contribution claim: A reproducible, fully-local pipeline that reads the public safeguarding record as structure; a taxonomy of measurement regimes and failure archetypes; a set of calibrated statistical models of the national outcome surface that assign a defensible charge probability by force and offence and quantify the postcode lottery; three reform simulations that put numbers and intervals on what complainant support, force-levelling and early recognition would each buy; a structural leverage index $L(v) = C_B(v)\, f(v)$ that ranks reform targets by combined brokerage and prevalence; and an empirical map from each recovered structure to a named reform owner. The central empirical finding is that institutional failure here is a problem of signal transmission and late recognition within an already dense network, dominated by institutional self-protection — and that its national shadow is an outcome surface that sheds most recorded sexual offences through victim withdrawal, unevenly enough across forces that where a case is reported measurably changes its odds.

Measured structure (evidence)	Reform it licenses	Owner / level
Recorded rape Gini $= 0.52$ across states; step-changes track legal reform	Harmonise definition, counting rule and survey mode before any cross-state ranking; publish regime metadata with every count	Statistical authorities; UNODC/Eurostat
Victim withdrawal is the modal outcome for recorded rape (49.7%); only 3.5% charged	Resource complainant support across the whole report-to-charge process; treat the withdrawal rate as the headline performance metric, not the charge rate alone	Police; Home Office (Operation Soteria)
Charge rate spans 2.8–11.0%; 18 of 44 forces outside the 99.8% funnel limits; place is the second-strongest driver	A national charging-performance standard with funnel-based monitoring; treat statistical outliers as automatic audit triggers	HMICFRS; CPS
External review is the central decision; absorption into external review equals absorption into prosecution	Move recognition upstream: independent, structured re-review triggered early, not only after catastrophe	Inspectorates; safeguarding partnerships
Dense 13-core, *zero* articulation points	Stop funding new connection infrastructure; fund signal quality and accountability at existing junctions	Government; commissioners
Highest leverage $L(v)$ at referral and investigation junctions	Place a named, accountable decision-owner at the referral/investigation broker points	Police; children's social care
An abuse-of-position or toxic-culture risk all but determines an oversight failure (mutual information ≈ 1 bit)	Auto-escalate oversight scrutiny whenever a position-of-authority or toxic-culture risk is recorded	Boards; CQC / Ofsted
Rule: reputation-over-protection $\Rightarrow$ allegation-not-acted-on (conf. 1.0)	Treat any recorded reputational consideration as a mandatory trigger to re-examine unacted allegations	Regulators; inspectorates
Dominant latent factor is institutional self-protection	Statutory duty to report with sanction for reputational suppression; remove discretion to protect the institution	Legislators (law reform)

Table 3: The structural map for reform. Each row anchors a reform to a measured structure recovered in this paper and assigns it to the level that could enact it. The reforms are offered as testable hypotheses: each predicts an observable change in the structures from which it was derived.

11 IMPLICATIONS

For **policing**, the actionable implication is concentrated at the referral and investigation junctions, which carry the highest structural leverage. The data argue for a named, accountable decision-owner at those points whose specific duty is to act on signal the system already holds — the abuse-of-position risk that near-deterministically precedes an oversight failure, the going-missing pattern that precedes an unacted allegation — rather than to forward it onward into an already saturated network.

For **policing and prosecutors at national scale**, the outcome models convert the postcode lottery into a management instrument. A calibrated per-force charge probability with confidence intervals turns "some forces charge more" into a defensible, monitorable standard: forces outside the funnel's control limits are audit triggers, not league-table entries, and the gap between a force's predicted and observed rate is a supervision signal. The dominance of victim withdrawal — half of all recorded rapes — argues that the headline performance metric should be the withdrawal rate, not the charge rate alone, because it measures the thing the system can most directly change: whether a complainant is supported through the process rather than lost from it. This is the empirical case for the complainant-support emphasis of the Operation Soteria model, made on the system's own outcome data.

For **law and legislators**, the dominant latent factor — institutional self-protection — is the target. The structural evidence supports a statutory duty to report with a real sanction for suppression in the service of reputation, because the recurring generator of failure is not that institutions could not see the harm but that protecting themselves competed with protecting the victim, and won. The association rule linking recorded reputational considerations to unacted allegations, at confidence one, is the empirical signature of that competition.

For **government and inspectorates**, the implication is a reallocation. Resources directed at new information-sharing infrastructure are, on this evidence, aimed at a problem that the network analysis says does not exist; the same resources directed at signal quality, at upstream independent re-review, and at accountability at the broker junctions are aimed at the problem that does. Inspection regimes can encode the association rules directly as triggers: when the visible antecedent appears, audit for the hidden consequent.

For **researchers**, the implication is methodological. The structural reading recovered patterns that a semantic reading of the same documents does not (the weak text–structure correlation makes this concrete), and it did so reproducibly. The approach generalises to any domain that accumulates a large narrative record of institutional failure — aviation, healthcare, financial regulation — where the same borrowed tools would license the same kind of structural map.

12 LIMITATIONS

The limitations are real and we state them plainly, because the credibility of a structural claim depends on the honesty of its bounds. The human-validated corpus is small ($n = 11$) and purposive, so the failure-level analytics — association rules, latent factors, mutual information — are exploratory structure, not estimates of population prevalence; a lift of 3.67 on eleven cases identifies a candidate relationship, not a confirmed rate. The scaled 61-case corpus is model-coded and not human-validated, which is precisely why its interpretive fields are not used for the failure analytics; its agency and pathway structure, where the model is reliable, is.

The Markov grammar is therefore estimated on model-coded decisions and should be read as the structure of the *documented* pathway, not of population attrition. The international comparison is illustrative and rests on one Eurostat extract. The outcome models carry their own bounds, which we state with equal candour. They are fit on aggregated force-by-offence cells for a single year (2023/24), so they are cross-sectional and ecological: the low pseudo-R^2 is honest evidence that most of the variation between individual offences turns on case facts the open data do not contain, and the force effect, though real and significant, is an association at the area level, not a demonstrated cause. The funnel assumes a common underlying rate that differences in case mix could partly explain; we adjust for offence subgroup but not for the full mix. "Victim does not support action" is the police-recorded outcome category, not a complainant's judgement, and we treat it strictly as the system's recorded reason for closure. The older and modern outcome frameworks are not directly comparable, so we do not splice them into one trend. The analysis is observational throughout: every arrow is a hypothesis, and the leverage index ranks structural pivotality, which is necessary but not sufficient for causal leverage. The cross-domain analogies, though stated as structural identities, are still heuristics that import each tool's own assumptions — independence in the lift calculation, linearity and non-negativity in the factorisation, the Markov property in the chain — and those assumptions are only approximately met. The reform simulations inherit every one of these bounds and add their own: they are what-if arithmetic, not forecasts, and their magnitudes move with stated priors — the charge-to-conviction rate, the fraction of withdrawals a supported process could recover, the reweighting parameter δ — which we set conservatively and print on the figures, but which only a prospective evaluation can pin down. They assume, too, that a force *could* move to another's rate without changing its case mix, which the postcode lottery's residual case-mix differences make an upper bound rather than a promise. Finally, a 1.2-billion-parameter local model is far weaker than a frontier model, and our reliability figures are a floor, not a verdict on what automated reading could achieve.

13 FUTURE RESEARCH

Five lines follow directly. First, the reliability ceiling can be raised by re-running the coding with a stronger local model and re-measuring κ, which would let the interpretive analytics move onto the full corpus. Second, the outcome models should be extended from one year to the full multi-year panel, which would let them trace the charge-rate collapse and its partial recovery and turn the cross-sectional postcode lottery into a trajectory; the staggered rollout of Operation Soteria across forces is a natural experiment that a difference-in-differences design could exploit to move from association to cause. Third, the early-warning rules are prospectively testable: an inspectorate could record the antecedent cues at the time of a case and measure whether the predicted consequent failures follow, turning a structural regularity into a validated instrument. Fourth, the leverage index invites a quasi-experimental test — where a broker junction has been made accountable by a local reform, the structures from which the index was derived should change measurably. Fifth, the whole apparatus should be ported to an adjacent failure record — healthcare never-events, aviation incident reports — both to test its generality and to strengthen the structural analogies by seeing where they break.

14 CONCLUSION

The public record of how institutions count, resolve and fail to prevent sexual offences has been read, until now, almost entirely as prose. Read instead as structure, with tools borrowed from sciences that have already solved structurally similar problems, it gives up a different and more actionable picture, and it does so at three scales that turn out to be one argument. The numbers that frame the field measure regime, not harm, and an inequality lens says exactly how much. The outcomes those numbers meet are steep and uneven: at national scale only one recorded rape in twenty-eight is charged, half are closed because the complainant withdraws, and where a case is reported changes its odds beyond chance — a surface we can model to a calibrated probability. And the failures behind that surface are not a formless catalogue: they come in predictable bundles, they are generated by three latent archetypes dominated by institutional self-protection, they follow a decision grammar whose recognition arrives late and often only when forced, and they expose their leverage at identifiable junctions inside a network that was never actually disconnected. The single most important finding is that last one: the institutions were connected and the signal was present; what failed was transmission and recognition, and the national outcome surface is what that failure looks like spread across two hundred thousand cases a year. That is a harder problem than fragmentation, but a more tractable one, because it can be located and measured — and this paper offers a map, anchored to measured structure and assigned to the levels that could act, for doing so. And because the repair is simulated rather than asserted, the map carries numbers: on stated, conservative assumptions, supporting complainants would roughly treble convictions per recorded rape, closing the postcode lottery would add thousands of charges a year, and moving recognition upstream would turn half of the cases that now end in a retrospective review into a timely prosecution. Three in a hundred is not a fact of nature; it is a property of a system, and a system can be changed. None of it requires, or permits, any judgement of the people who reported the harm; the reading is of the record, and of the system that made it.

Ethics and scope. This study analyses only public, already-anonymised documents and official statistics. It performs record-integrity analysis — how the institutional record was produced and where it failed — and at no point scores or models the credibility of a person who reported harm, which is out of scope and excluded by design. The published dataset carries derived structural codes and provenance only; for the scaled corpus no report text is redistributed, respecting the publishers' copyright and any reporting restrictions. Findings are descriptive and associational; causal language is confined to explicitly labelled hypotheses.

REFERENCES

Agrawal, R., T. Imieliński, and A. Swami (1993). "Mining Association Rules Between Sets of Items in Large Databases". In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216.

Brandon, M., S. Bailey, P. Belderson, R. Gardner, P. Sidebotham, J. Dodsworth, C. Warren, and J. Black (2009). *Understanding Serious Case Reviews and Their Impact: A Biennial Analysis of Serious Case Reviews 2005–07*. Tech. rep. DCSF-RR129. London: Department for Children, Schools and Families. URL: https://www.gov.uk/government/publications/understanding-serious-case-reviews-and-their-impact.

Crowther, T. (2022). *Independent Inquiry: Telford Child Sexual Exploitation*. Tech. rep. Telford and Wrekin Council. URL: https://www.telford.gov.uk/.

Eurostat (2026). *EU Sexual Violence and Rape Offences Up in Last 10 Years*. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20260429-2.

Freeman, L. C. (1977). "A Set of Measures of Centrality Based on Betweenness". In: *Sociometry* 40.1, pp. 35–41.

Gini, C. (1912). "Variabilità e Mutabilità". In: Contribution to the study of distributions and relations of statistical series. Bologna: Tipografia di Paolo Cuppini.

Her Majesty's Inspectorate of Constabulary (2014). *Crime-recording: Making the Victim Count*. Tech. rep. HMIC. URL: https://www.justiceinspectorates.gov.uk/hmicfrs/publications/crime-recording-making-the-victim-count/.

Home Office (2023). *Operation Soteria Year One Report*. Tech. rep. Home Office. URL: https://www.gov.uk/government/publications/operation-soteria-year-one-report.

— (2025). *Crime Outcomes in England and Wales 2024 to 2025*. Tech. rep. Home Office. URL: https://www.gov.uk/government/statistics/crime-outcomes-in-england-and-wales-2024-to-2025.

— (2026). *Counting Rules for Recorded Crime*. URL: https://www.gov.uk/government/publications/counting-rules-for-recorded-crime.

Independent Inquiry into Child Sexual Abuse (2022). *The Report of the Independent Inquiry into Child Sexual Abuse*. Tech. rep. IICSA. URL: https://www.iicsa.org.uk/reports-recommendations/publications/inquiry/final-report.html.

Jay, A. (2014). *Independent Inquiry into Child Sexual Exploitation in Rotherham 1997–2013*. Tech. rep. Rotherham Metropolitan Borough Council. URL: https://www.rotherham.gov.uk/.

Kelly, L., J. Lovett, and L. Regan (2005). *A Gap or a Chasm? Attrition in Reported Rape Cases*. Tech. rep. Home Office Research Study 293. London: Home Office. URL: https://cwasu.org/resource/a-gap-or-a-chasm-attrition-in-reported-rape-cases/.

Kemeny, J. G. and J. L. Snell (1976). *Finite Markov Chains*. New York: Springer-Verlag.

Lee, D. D. and H. S. Seung (1999). "Learning the Parts of Objects by Non-negative Matrix Factorization". In: *Nature* 401.6755, pp. 788–791.

Lisak, D., L. Gardner, S. C. Nicksa, and A. M. Cote (2010). "False Allegations of Sexual Assault: An Analysis of Ten Years of Reported Cases". In: *Violence Against Women* 16.12, pp. 1318–1334. URL: https://journals.sagepub.com/doi/10.1177/1077801210387747.

Morselli, C. (2009). *Inside Criminal Networks*. New York: Springer.

Office for National Statistics (2021). *Nature of Sexual Assault by Rape or Penetration, England and Wales: Year Ending March 2020*. Tech. rep. Office for National Statistics. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/natureofsexualassaultbyrapeorpenetrationenglandandwales/yearendingmarch2020.

Seidman, S. B. (1983). "Network Structure and Minimum Degree". In: *Social Networks* 5.3, pp. 269–287.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423.

Sidebotham, P., M. Brandon, S. Bailey, P. Belderson, J. Garstang, E. Harrison, A. Retzer, and P. Sorensen (2016). *Pathways to Harm, Pathways to Protection: A Triennial Analysis of Serious Case Reviews 2011 to 2014*. Tech. rep. London: Department for Education. URL: https://www.gov.uk/government/publications/analysis-of-serious-case-reviews-2011-to-2014.

Spiegelhalter, D. J. (2005). “Funnel Plots for Comparing Institutional Performance”. In: *Statistics in Medicine* 24.8, pp. 1185–1202.

United Nations Office on Drugs and Crime (2024). *Compiling and Comparing International Crime Statistics*. URL: https://www.unodc.org/unodc/en/data-and-analysis/Compiling-and-comparing-International-Crime-Statistics.html.

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang (2024). “Can Large Language Models Transform Computational Social Science?” In: *Computational Linguistics* 50.1, pp. 237–291. URL: https://direct.mit.edu/coli/article/50/1/237/118498.

A CONTROLLED VOCABULARIES AND METHODS

The coding scheme fixes four controlled vocabularies: *agencies* (21 terms, e.g. police, children’s social care, health, education, inspectorate, safeguarding partnership), *risk indicators* (18 terms, e.g. abuse of position or authority, organisational-culture risk, residential setting, going missing, online facilitation), *decision types* (15 ordered terms, e.g. referral made, investigation launched, strategy meeting, child-protection plan, prosecution/conviction, external review commissioned), and *record-integrity failure flags* (15 terms, e.g. allegation not acted on, reputation over protection, oversight inadequate, information not shared, crime not recorded). The flag entropy across the validated corpus is 13.0 bits, confirming that failures are diverse rather than concentrated on a single mode. Full definitions and the method formulas of Section 6 ship with the released code.

B REFERENCE CORPUS

ID	Case (public report)	Type	Year	Jurisdiction
CASE-01	Nottinghamshire	IICSA	2019	England & Wales
CASE-02	Lambeth	IICSA	2021	England & Wales
CASE-03	Custodial estate	IICSA	2019	England & Wales
CASE-04	Anglican / Ball	IICSA	2019	England & Wales
CASE-05	Religious settings	IICSA	2021	England & Wales
CASE-06	Oxfordshire CSE	SCR	2015	England & Wales
CASE-07	Rotherham	inquiry	2014	England & Wales
CASE-08	Telford	inquiry	2022	England & Wales
CASE-09	CSA in the family	CSPR	2024	England
CASE-10	Bexley (DHR)	DHR	2020	England & Wales
CASE-11	Residential schools	IICSA	2022	England & Wales

Table 4: The eleven human-validated cases. Institutions are named in the public reports; only victims are anonymised. The set spans review types, institutional settings and two decades by design.

C THE OUTCOME DATA AND MODELS

The outcome arm draws on three Home Office open-data releases: the modern outcomes open data for 2023/24 ($\approx 712{,}000$ rows, force $\times$ offence $\times$ quarter $\times$ outcome), the 2005/06–2013/14 sanctions series ($\approx 287{,}000$ rows), and Community-Safety-Partnership recorded volumes for the decade to 2024 (≈ 1.05 million rows across 359 local areas) — roughly two million rows in total, fetched into the git-ignored raw directory, from which only aggregated

modelling tables and provenance are committed. The parser maps the twelve-group outcome taxonomy to analytic categories and recovers the recorded-offence denominator by summation. Models are fitted on weighted cells: a binomial GLM (`statsmodels`) for the charge and withdrawal probabilities and their odds ratios, nested GLMs for the deviance decomposition, exact binomial funnel limits for the institutional comparison, a Theil index for between-force inequality, and a histogram gradient- boosting classifier (`scikit-learn`) with permutation importance and a calibration check for the predictive surface.

D REPRODUCIBILITY

All figures and tables regenerate offline from committed derived data via the released pipeline (`make reproduce`). Raw inputs are git-ignored; derived structural codes, the outcome modelling tables, provenance and the forensics and model result sets are committed. The outcome models live in the `outcomes`, `models` and `models_figures` modules; the cross-domain structural analytics in `forensics` and `forensics_figures`; the measurement analysis reads a committed Eurostat extract. The local model is `liquid/lfm2.5-1.2b` served through an OpenAI-compatible endpoint, with embeddings from a local `nomic-embed-text` model; chat extraction is decomposed per field and constrained to the controlled vocabulary by a JSON-schema grammar.