

THE DIGITAL POORHOUSE REVISITED

*Algorithmic Governance in Welfare Systems:
Harms, Governance Gaps, and an Evaluation Framework for Safer Deployment*

A Policy Brief

Nina Pham | The Field Center | University of Pennsylvania School of Policy and Practice

Welfare AI systems are producing measurable harms and governance has not caught up

1M+

benefit denials in Indiana
in 3 years

Eubanks, 2018

93%

error rate in Michigan's
MiDAS fraud system

MI Auditor General, 2016

\$1.8B

AUD settlement for
Australia's Robodebt

Holmes, 2023

The brief synthesizes evidence from seven systems across four jurisdictions to show that welfare AI harms stem from structural governance failures, not primarily from model accuracy, and proposes a five-principle evaluation framework for safer deployment.

Anchor claim: AI in welfare tends to function as governance-by-surveillance and procedural punishment unless constrained by enforceable rights, meaningful transparency, and benefit-maximizing design.

Three structural mechanisms convert welfare automation into measurable harm



Administrative friction becomes punishment

Rigid workflows convert minor documentation gaps into mass denials and penalties at scale impossible under manual review.



Outcome variables measure system behavior, not need

Models predict future government contact: investigations, placements, overpayments, rather than underlying conditions.



Opacity weakens accountability and rights

Scoring behind the scenes forecloses notice, explanation, contestation, and correction that welfare law requires.

The "digital poorhouse" framework explains why these harms are structural, not incidental

From social rights to conditional access

Eligibility becomes a test of bureaucratic compliance: fax deadlines, portal logins, document uploads not substantive need.

From public debate to administrative inevitability

Political choices about who deserves help get encoded as technical parameters and shielded from democratic contestation.

From human judgment to proxy decisions

Systems rely on administrative data legibility, entrenching surveillance patterns and foreclosing corrective action.

Modern ML systems do not escape the digital poorhouse critique, they deepen it

Measurement problem persists

Better prediction of a proxy outcome (e.g., future foster placement) deepens inequality if the outcome itself reflects institutional behavior.

Error costs scale

A single design flaw produces errors at population level; correction burden shifts to affected individuals (MiDAS: years of litigation).

Opacity is a rights problem

The SyRI court found that transparency without contestation power is compliance documentation, not accountability.

Transparency ≠ power shift

New disclosure frameworks (EU AI Act, UK ATR) document systems but may legitimate deployment without shifting power to claimants.

One spine question and two supporting questions organize the analysis

SPINE

What governance conditions determine whether welfare AI produces administrative harm versus public value?

SQ1: Measurement Validity

What do welfare models actually predict, and how often are they predicting system behavior rather than underlying need or harm?

SQ2: Due Process & Error Correction

What happens to appeals, recourse, and correction when decisions are automated and who bears the cost of errors?

Crosscutting: disparate impact (race, disability, SES); privacy and proportionality; comparative lessons from jurisdictions that constrained welfare algorithms.

Seven cases across four jurisdictions reveal consistent governance failures

System	Type	Key Harm	Governance Response	Status
Indiana Eligibility	Eligibility automation	1M+ denials; 54% spike yr 1	Contract cancelled; litigation	Terminated
MiDAS (Michigan)	Fraud detection	93% error rate; 40K false accusations	\$20M settlement; legislation	Replaced
AFST (Allegheny Co.)	Risk scoring	Racial disparities; proxy reliance	ACLU analysis; ongoing refinement	Active
VI-SPDAT (US)	Triage	Unvalidated scoring; disclosure bias	Discontinued by OrgCode 2023	Phasing out
SyRI (Netherlands)	Fraud/risk indication	Privacy violation; neighborhood targeting	Struck down by court (2020)	Terminated
Robodebt (Australia)	Debt recovery	526K debt notices; documented suicides	Royal Commission; \$1.8B settlement	Terminated
UC Advances (UK)	Fraud/risk scoring	Disparate impact concerns	ATR published; fairness assessment	Active

Administrative friction at scale: Indiana and Michigan automated punishment for procedural noncompliance

INDIANA (2006-2009)

54% denial spike

in first year of automated eligibility

1M+ denials in 3 years

driven by "failure to cooperate" codes

Three-document fax rule

and rigid deadline enforcement replaced caseworker flexibility

MIDAS, MICHIGAN (2013-2015)

93% error rate

in 22,000 reviewed fraud determinations

400% automatic penalties

applied without human review

\$20M class action settlement

(Bauserman v. UIA, 2024) after 9 yrs

Proxy measurement produces racial disparities even when models are technically accurate

AFST (ALLEGHENY COUNTY)

"Ever-in" features

(jail, Medicaid, juvenile probation) produce racially disparate scores

Removing them: AUC drops 0.002

negligible accuracy cost, meaningful equity gain

Human overrides reduced disparity

from ~20% to 9% but this safeguard is informal and unaudited

VI-SPDAT (NATIONWIDE)

Structured questionnaire

privileges specific vulnerability narratives

Penalizes people who cannot

navigate assessment or disclose trauma strategically

Discontinued by developer (2023)

citing validity and equity concerns; many jurisdictions still using it

Opacity at scale: SyRI and Robodebt show that undisclosed systems produce rights violations

SyRI (NETHERLANDS)

17 categories of government data
cross-referenced through undisclosed model

Deployed exclusively
in low-income and minority neighborhoods

Struck down (Feb 2020)
under ECHR Art. 8: opacity prevented
proportionality review

ROBODEBT (AUSTRALIA)

526,000 debt notices
generated by legally deficient income averaging

Reversed burden of proof
recipients had to disprove debts years later

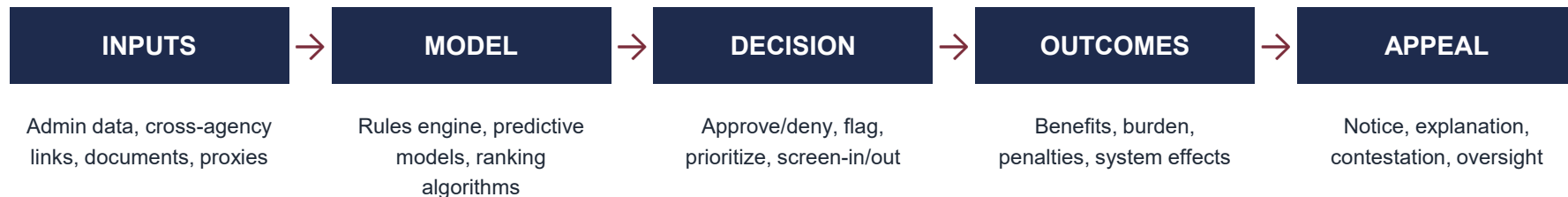
\$1.8B AUD class action settlement
Royal Commission (2023); documented suicides

Across all seven cases, governance failure not model failure, is the primary driver of harm

Of seven systems examined, five have been terminated, replaced, or struck down by courts. In every case, the system operated for years before governance caught up.

- 1 No system was shut down because of a model accuracy problem.
- 2 Every termination or legal challenge was driven by due process, privacy, or proportionality failures.
- 3 External pressure (litigation, media, advocacy) was required in every case—internal oversight was insufficient.
- 4 Human-in-the-loop safeguards, where they existed, were informal and uncoded.

The welfare AI decision pipeline has five stages and two recurring failure traps



⚠️ PROXY TRAP

Concentrates at Stage 2: when outcome variables measure system contact rather than need, and features encode prior surveillance, the model optimizes prediction while deepening inequality.

⚠️ OPACITY TRAP

Concentrates at Stage 5: when the person cannot see, understand, or contest what drove the decision, and oversight bodies lack access to model documentation.

The evaluation framework assesses welfare AI across five governance dimensions

- | | | |
|----------|--|---|
| 1 | Measurement Validity | Does the model predict need/harm, or system behavior? |
| 2 | Due Process | Can affected persons understand, contest, and correct decisions? |
| 3 | Transparency & Contestability | Can oversight bodies and communities scrutinize the system? |
| 4 | Equity & Non-Discrimination | Does the system produce disparate outcomes by race, disability, income? |
| 5 | Proportionality & Data Minimization | Is data integration proportionate to the legitimate aim? |

Five design principles translate case evidence into actionable policy safeguards

1

Pre-deployment impact assessments

with affected community participation

Robodebt Rec. 18; AI Now AIA

2

Pre-deprivation notice, explanation, and human review

for all automated determinations affecting benefits

Bauserman v. UIA; SyRI ruling

3

Audit outcome variables and proxy features

for measurement validity and disparate impact

AFST "ever-in" analysis; NIST AI RMF

4

Error-cost asymmetry as a design constraint

minimize false positives; cap automated penalties

MiDAS 400% penalty; Kohl, 2024

5

Sunset clauses and mandatory independent evaluation

with affirmative reauthorization required

MiDAS 2-yr delay; Robodebt 4-yr persistence

Error-cost asymmetry is the most underused safeguard: false positives in welfare are catastrophic

FALSE POSITIVE COSTS (to individuals)

- Wrongful denial of benefits
- Wrongful fraud accusation + automatic penalties
- Wrongful investigation (child welfare)
- Wage garnishment, credit destruction
- Bankruptcy, housing loss
- Years of litigation to correct

FALSE NEGATIVE COSTS (to system)

- Missed fraud (diffuse fiscal cost)
- Missed risk (diffuse system cost)
- Typically reversible on detection
- Does not destroy individual livelihood

Design implication: Choose thresholds that minimize false positives. Eliminate or cap automatic penalties for algorithmically flagged determinations.

The framework is a pre-deployment and ongoing assessment tool, not a one-time compliance checklist

1

Complete before deployment

Assess all five governance dimensions at all five pipeline stages with affected community input.

2

Update at defined intervals

At minimum annually and after any model retraining or scope expansion.

3

Include external voices

Agency self-assessment alone is insufficient—Indiana, Michigan, and Australia all had internal processes that failed.

4

Make assessments public

Publish completed assessments so advocates, researchers, and oversight bodies can scrutinize them.

The question is not whether welfare AI "works", it is whether governance has caught up to the harms it produces

The evidence is clear: welfare AI systems can produce measurable, large-scale harm even when they function as technically designed. The problem is governance, not accuracy.

- 1 Mandate pre-deployment impact assessments with affected community participation.
- 2 Require pre-deprivation notice, explanation, and human review for automated welfare determinations.
- 3 Audit outcome variables for measurement validity and disparate impact.
- 4 Design for error-cost asymmetry: welfare false positives are catastrophic.
- 5 Build sunset clauses and mandatory independent evaluation into every deployment.

These are not aspirational recommendations. They are the minimum conditions under which welfare AI can be deployed without reproducing the governance failures documented across seven systems and four jurisdictions.

DISCUSSION

Questions, Challenges, and Next Steps