

# Gradual Disempowerment at Scale: Measuring Cumulative Agency Erosion in Algorithmic Management of Indian Delivery Workers

Pavitra, PaXus Labs / IIT Kanpur Fellow | Apart Research Global South AI Safety Hackathon 2026

Regional track: Asia | Sub-track: Socio-Economic Impacts, Algorithmic Management, Delivery Platforms and Gig Work

**Abstract.** AI safety has many tools for acute model failures but few evals for gradual disempowerment: incremental loss of human influence as automated systems control decision channels. We operationalize Kulveit et al. 2025 as a five-dimension Gradual Disempowerment Score (GDS) for Indian delivery platforms. The pipeline combines Fairwork India evidence for Swiggy and Zomato from 2021-2024, NITI Aayog gig-workforce estimates, and a clearly labeled synthetic predictive lock-in simulation where worker-level allocation logs are unavailable. Results are mixed. GDS peaks around 2022-2023, then falls in 2024 as Fairwork-evidenced contracts and management improve. In 2024, Swiggy scores 37.88 and Zomato 38.28 on a 0-100 risk scale. GDS is best read as a replicable diagnostic for hidden agency erosion, not proof that every public indicator worsened.

## 1. Introduction

Gradual disempowerment is a safety failure mode in which systems become less dependent on human labor, judgment, and feedback, causing human influence over economic and social outcomes to erode without a single dramatic takeover event. Kulveit et al. argue that this can be catastrophic because explicit control channels and implicit reliance on human participation can both weaken over time.

India's app-based delivery sector is a useful near-term domain for this frame. Swiggy and Zomato workers are managed through allocation, routing, rating, penalty, incentive, and grievance systems. These systems need not be malicious for agency erosion to occur. A platform can improve one procedural safeguard while other channels, such as worker voice or rating-based opportunity, remain structurally weak.

This project makes three contributions: (1) a formal five-dimension GDS metric for cumulative agency erosion, (2) a runnable open-source pipeline computing Swiggy and Zomato scores from 2021-2024, and (3) sensitivity analysis showing where rankings are robust and where they depend on weight choices.

## 2. Related Work

**Gradual disempowerment.** The prompt referenced Kulveit et al. 2021, but the source verified during this sprint is Kulveit et al. 2025 on arXiv. We cite the verified source rather than inventing a 2021 citation. The paper motivates measuring erosion of human influence across economic, cultural, and state systems.

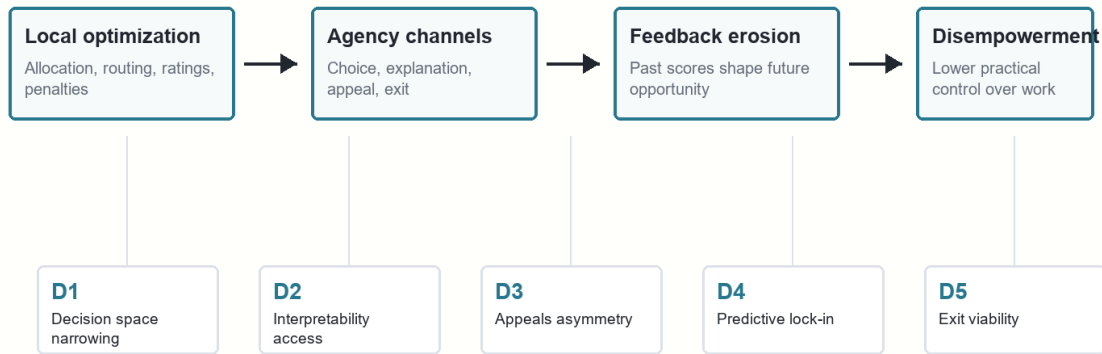
**Algorithmic management.** Prior work on algorithmic management shows that platforms can transfer managerial functions to ranking, monitoring, ratings, nudges, and automated discipline. Lee et al. 2015 introduced key worker-facing concerns around algorithmic and data-driven management. Recent India-specific work documents invisible labor, repetitive app interactions, gamification, and opacity in food delivery work.

**Gap.** Labor studies measure pay, conditions, classification, and fairness. AI safety evals usually measure model behavior. What is missing is a bridge metric that asks whether a deployed algorithmic management system is cumulatively reducing worker agency in the Kulveit sense. GDS fills that gap as an auditable first pass.

## 3. Methodology

## Figure 4. Mapping GDS to gradual disempowerment

The eval tracks loss of worker influence before a sudden failure is visible.



**GDS = weighted sum of D1-D5, with robustness checked by weight perturbation.**

Figure 4. Conceptual mapping from local platform optimization to cumulative loss of worker influence. The five GDS dimensions are designed to be measurable from public or auditable platform evidence.

| Dimension           | Proxy                                         | Source                  | Direction                                             |
|---------------------|-----------------------------------------------|-------------------------|-------------------------------------------------------|
| D1 Decision space   | Representation, management, and contract risk | Fairwork principles     | Higher risk means fewer practical choices             |
| D2 Interpretability | Inverse Fair Contracts score                  | Fairwork principle 3    | Higher risk means less understandable work terms      |
| D3 Appeals          | Inverse Fair Management score                 | Fairwork principle 4    | Higher risk means weaker due process                  |
| D4 Lock-in          | Synthetic rating-to-allocation risk           | Simulation              | Higher risk means past scores constrain future access |
| D5 Exit viability   | Gig workforce growth plus pay risk            | NITI Aayog and Fairwork | Higher risk means weaker outside option               |

For each platform-year, each dimension is scored from 0 to 100, where 100 is maximum disempowerment risk. Baseline GDS is the weighted sum of D1-D5 with weights D1=0.25, D2=0.20, D3=0.20, D4=0.20, and D5=0.15. D1 receives the largest weight because narrowing the worker's actionable choice set is the most proximate loss-of-control channel. D2-D4 are weighted equally because explanation, recourse, and future opportunity are mutually reinforcing control loops. D5 is weighted lower because exit viability is important but less platform-specific in the available public data.

Fairwork evidence is used for contract transparency, management due process, worker representation, pay, and safety. Fairwork's methodology combines desk research, worker interviews, and platform evidence, with points awarded only when clear evidence exists. NITI Aayog's 2022 report supplies national gig workforce scale: 7.7 million workers in 2020-21 and a projection of 23.5 million by 2029-30. D4 is synthetic and is not treated as observed evidence.

Sensitivity analysis perturbs each baseline weight independently by +/-10 percent, renormalizes the weights, and repeats the ranking over 1,000 samples. Stability is reported as the share of samples that preserve the baseline top-risk platform for a given year.

4. Results

| Platform | 2021  | 2022  | 2023  | 2024  |
|----------|-------|-------|-------|-------|
| Swiggy   | 34.0  | 48.38 | 49.86 | 37.88 |
| Zomato   | 47.73 | 48.78 | 50.26 | 38.28 |

Table 1. Baseline GDS scores. Higher values indicate higher measured agency erosion.

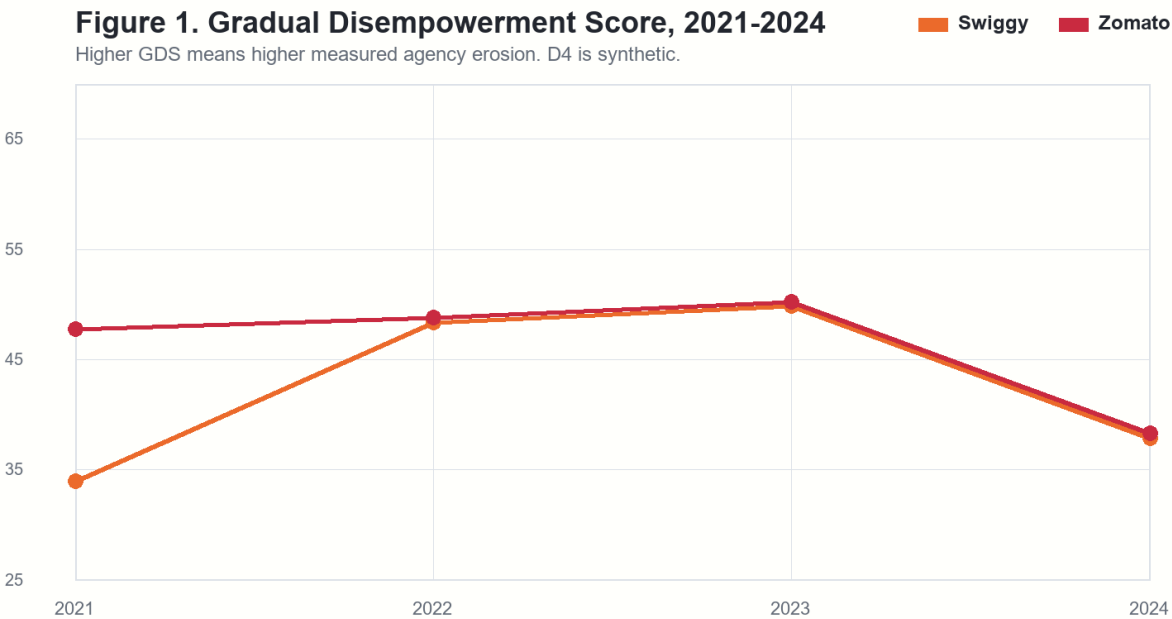


Figure 1. GDS trajectories for Swiggy and Zomato. The public-data composite peaks in 2022-2023 and falls in 2024, mainly because Fairwork-evidenced contracts and management improve.

**Figure 2. Dimension profile in 2024**  
Each axis is a 0-100 risk score. Higher means worse agency erosion.

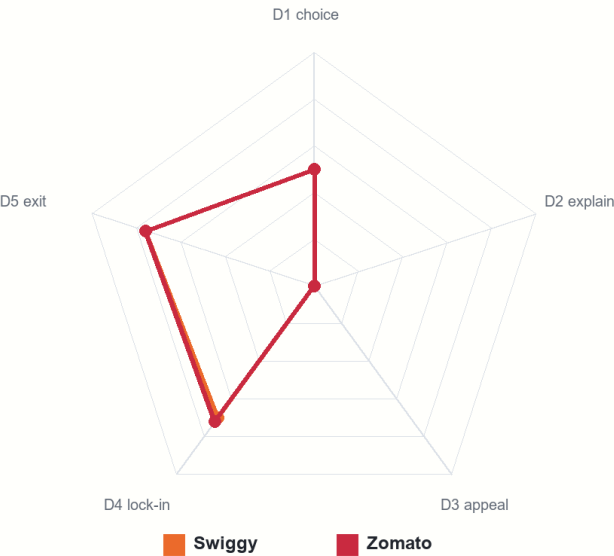


Figure 2. In 2024, both platforms have low D2 and D3 risk in the Fairwork-based proxy, but remain high on D5 and moderate on D1 and D4.

### Figure 3. Weight sensitivity under +/-10 percent perturbation

Horizontal bands show sampled GDS ranges after weight renormalization.

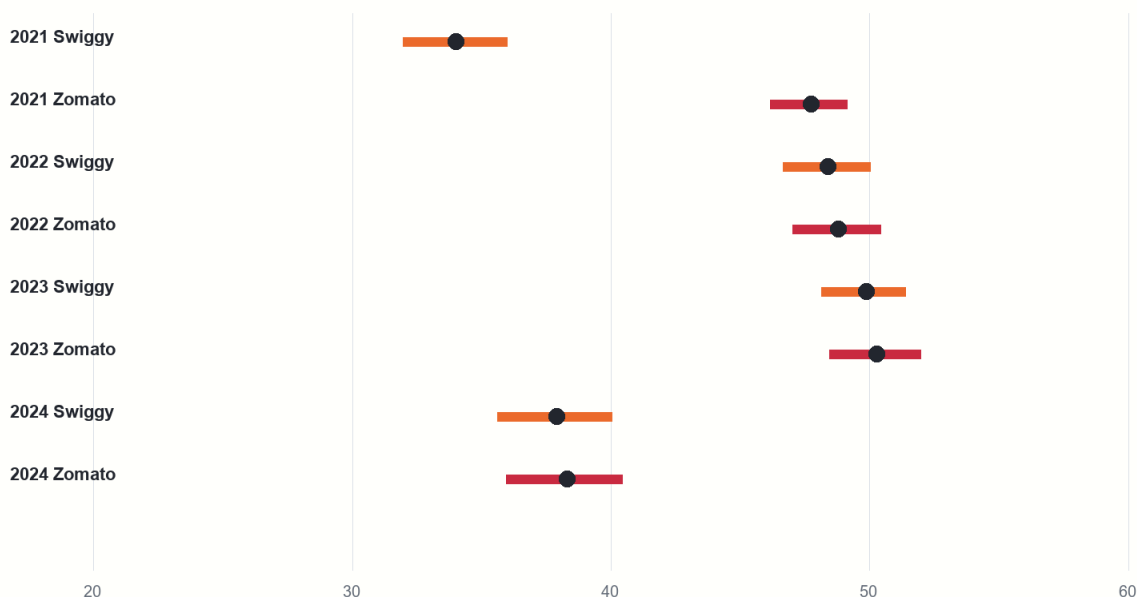


Figure 3. Weight sensitivity ranges under +/-10 percent perturbation. The 2023 rank is less stable because scores are close and D2/D3 move differently across platforms.

Zomato is the baseline higher-risk platform in all four years, but the difference is substantively small by 2024. Rank stability by year is 2021: 1.0, 2022: 1.0, 2023: 0.67, 2024: 1.0. This means 2021, 2022, and 2024 rankings are robust under this perturbation protocol, while 2023 is only moderately stable.

The strongest result is not that disempowerment monotonically increases. It does not in this public-data implementation. The stronger claim is that a measurable eval can expose which channels improve, which remain weak, and which require better data. In this run, public evidence suggests procedural improvement on contracts and due process, persistent weakness on representation and pay, and unknown but potentially important predictive lock-in.

## 5. Discussion

The immediate policy implication is modest but concrete: AI safety governance should not wait for internal platform logs before defining disclosure standards. India-specific AI and platform-labor governance could require auditable explanation, appeal, representation, and allocation-impact evidence. The same eval structure could also be adapted to credit scoring, welfare eligibility, healthcare triage, and other domains where automated decisions narrow agency gradually rather than causing a single visible failure.

## 6. Limitations and Dual-Use Considerations

**Data limitations.** D4 is synthetic because platforms do not publish worker-level histories linking ratings, acceptance, cancellations, and future order volume. D5 uses national gig workforce growth, not platform-specific income dependency. Fairwork scores are rigorous but not representative surveys; they are evidence thresholds based on desk research, worker interviews, and platform evidence. Cross-year comparison assumes Fairwork scoring remains stable enough for longitudinal use.

**Construct limitations.** GDS is a new construct. It has not yet been externally validated against worker-reported agency, health, income stability, or ability to contest platform decisions. A field validation should ask workers whether GDS changes track perceived control.

**Dual-use.** A platform could misuse a GDS-style scorecard for compliance theater, optimizing visible documentation while leaving actual allocation power opaque. To reduce this risk, future versions should require worker-side evidence, audit trails for automated decisions, and publication of uncertainty intervals, not just a single score.

**Future work.** The next sprint should collect 5-10 structured Bengaluru worker interviews, extend the eval to Rapido, Dunzo, Urban Company, and quick-commerce platforms, and validate D4 with opt-in worker app traces or platform disclosures.

## 7. Conclusion

GDS is a first operational eval for gradual disempowerment in algorithmic labor management. Applied to Swiggy and Zomato from 2021-2024, it gives a mixed but useful result: public evidence shows improvement in some procedural safeguards, persistent weakness in worker voice and pay, and unresolved risk from predictive lock-in. That is exactly why this category needs measurement. Without an eval, gradual disempowerment remains a compelling story; with GDS, it becomes a hypothesis that can be tested, challenged, and improved.

## Code and Data

Code artifact: local folder ``gig_safety_eval`` in this submission workspace. Generated outputs are in ``gig_safety_eval/results``, ``gig_safety_eval/paper/figures``, and ``output/pdf``.

Data: Fairwork India ratings and platform pages, Fairwork methodology, NITI Aayog 2022 report, source catalog in ``data/ilo_survey_data.csv``, and synthetic D4 data documented in ``data/synthetic_README.md``.

## References

1. Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. 2025. Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. arXiv:2501.16946. <https://arxiv.org/abs/2501.16946>
2. Fairwork. 2024. Fairwork India Ratings 2024: Labour Standards in the Platform Economy. <https://fair.work/en/fw/publications/fairwork-india-ratings-2024-labour-standards-in-the-platform-economy/>
3. Fairwork. Methodology. <https://fair.work/en/fw/methodology/>
4. Fairwork. India ratings, 2021-2024. <https://fair.work/en/ratings/india/>
5. NITI Aayog. 2022. India's Booming Gig and Platform Economy: Perspectives and Recommendations on the Future of Work. [https://www.niti.gov.in/sites/default/files/2022-06/25th\\_June\\_Final\\_Report\\_27062022.pdf](https://www.niti.gov.in/sites/default/files/2022-06/25th_June_Final_Report_27062022.pdf)
6. Lee, M. K., Kusbit, D., Metsky, E., and Dabbish, L. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. CHI 2015. <https://doi.org/10.1145/2702123.2702548>
7. Suvarnapathaki, P., Shah, V., Negi, S., and Rangaswamy, N. 2025. The Boring and the Tedious: Invisible Labour in India's Gig-Economy. arXiv:2504.17697. <https://arxiv.org/abs/2504.17697>

## LLM Usage Statement

We used Codex to assist with research synthesis, code scaffolding, figure generation, and draft writing. All empirical results in this report were computed by the included pipeline. Claims are limited to verified public sources and explicitly labeled synthetic assumptions.