

Consent at Population Scale

Governing public data for AI in India

Published

July 2026

Author

Rahul Rai

Series

AI as Public Infrastructure, No. 2

Key findings

India's digital public infrastructure has produced one of the world's richest records of how a developing economy works. Using that data to build AI suited to Indian languages and conditions is now official policy direction. The legal and institutional framework for doing so safely does not yet exist, and India's own evidence shows how data systems built for efficiency can exclude the people they are meant to serve.

16,077

Datasets listed on AIKosh, up from about 300 at launch in March 2025

0

Mentions of "artificial intelligence" in the DPDP Act and Rules

May 2027

When most substantive DPDP provisions take effect

- **The law does not address model training.** The DPDP Act's research exemption covers processing not used for decisions about individuals. It does not say whether, or when, training an AI model qualifies.
- **Research becomes deployment.** A model trained for research is valuable precisely because it can later be used on individuals. Nobody is assigned to check when that line is crossed.
- **Nobody grants or supervises access.** Each data holder decides alone, with no permit system, no public register and no common standard.
- **India has the technology; Europe has the institutions.** India's DEPA for Training keeps sensitive data in confidential clean rooms. The EU's Health Data Space and Finland's Findata show how independent access bodies issue and supervise permits.
- **The main risk is exclusion, not leakage.** A rigorous Jharkhand evaluation found that stricter biometric authentication raised beneficiaries' costs and cut benefits for those with incomplete records. AI trained on administrative data and used for targeting could repeat that pattern at greater scale.
- **Recommendation:** clarify the law on training; create independent data access bodies, starting with health and agriculture; make compute-to-data the default; draw a firm line between training and decisions about individuals; publish a register of access; and share the benefits of models trained on public data.

Contents

Executive summary	4
SECTION 1 · The opportunity and its limits	5
Why public data matters for Indian AI	5
AIKosh: volume is not value	5
SECTION 2 · The legal gap	7
What the DPDP framework says	7
Three problems	7
SECTION 3 · Models that work	9
DEPA for Training (India)	9
The European Health Data Space	9
Findata (Finland)	9
SECTION 4 · The critics are partly right	11
Evidence from Jharkhand	11
The lesson for AI	11
Other objections that deserve weight	12
SECTION 5 · A tiered model for access	13
SECTION 6 · Six proposals	14
1. Clarify the law on model training	14
2. Create independent data access bodies	14
3. Make compute-to-data the default	14
4. Draw a bright line between training and decisions	14
5. Publish a public register	15
6. Share the benefits	15
SECTION 7 · Open questions	16
SECTION 8 · Conclusion	17
About this report	17
Sources	18

Executive summary

AI models trained mostly on English-language data from wealthy countries perform worse on Indian languages, diseases, soils and informal economic patterns. India's public data, from health encounters to farm records to welfare transactions, could close that gap. The India AI Governance Guidelines endorse this direction, calling for DEPA-style consent architectures to support responsible AI training.

Three gaps stand in the way. **The law is silent on AI.** The Digital Personal Data Protection (DPDP) Act's research exemption was not written with model training in mind. **Access is unmanaged.** Every ministry, state department and hospital network decides for itself who may use its data and on what terms. **Uses are not separated.** Nothing stops a model trained under a research exemption from later being used to score, flag or exclude individuals.

The stakes are concrete. India's best evidence on data-driven welfare comes from a large randomised evaluation in Jharkhand, which found that stricter authentication raised costs for beneficiaries and reduced benefits for those whose records were incomplete. AI-based targeting systems, such as the vulnerability scoring proposed in Andhra Pradesh, would rely on the same kinds of records.

This report argues for a governed middle path between locking public data away and opening it up. India already has the technical building blocks in DEPA for Training. What it lacks are the institutions: an independent body to grant and supervise access, and clear rules on what models trained on public data may be used for. **The DPDP Act's substantive provisions take effect in May 2027. Building these institutions before then is the opportunity.**

Recommendations at a glance

1. **Clarify the law on model training** through MeitY guidance or an amendment to the DPDP Rules.
2. **Create independent data access bodies**, starting with health and agriculture, separate from the departments that use the data.
3. **Make compute-to-data the default** for sensitive public data, using clean-room environments with logged, auditable runs.
4. **Draw a bright line between training and decisions.** Models trained under the research exemption require a separate assessment, group-level error testing, human review and appeal before use on individuals.
5. **Publish a public register** of every access permit, its purpose, user and results.
6. **Share the benefits** through public-interest licence terms on models trained on public data.

SECTION 1

The opportunity and its limits

Why public data matters for Indian AI

General-purpose AI models learn from what is available online, which is overwhelmingly English and drawn from high-income countries. They know less about crop diseases in Vidarbha, the clinical presentation of tuberculosis in Indian primary care, or the cash flows of a kirana store. India's digital public infrastructure has generated exactly this kind of data, at population scale, under public authority. Used well, it could make Indian AI more accurate for Indian conditions and less dependent on foreign models.

Policy has moved in this direction. The India AI Governance Guidelines of November 2025 call for leveraging digital public infrastructure and for “DEPA-style consent architectures to support responsible AI training.”¹

AIKosh: volume is not value

The IndiaAI Mission's dataset platform, AIKosh, has grown rapidly. It listed about 300 datasets at launch in March 2025,² 7,541 by February 2026³ and more than 16,000 datasets and 372 models by October 2026.⁴

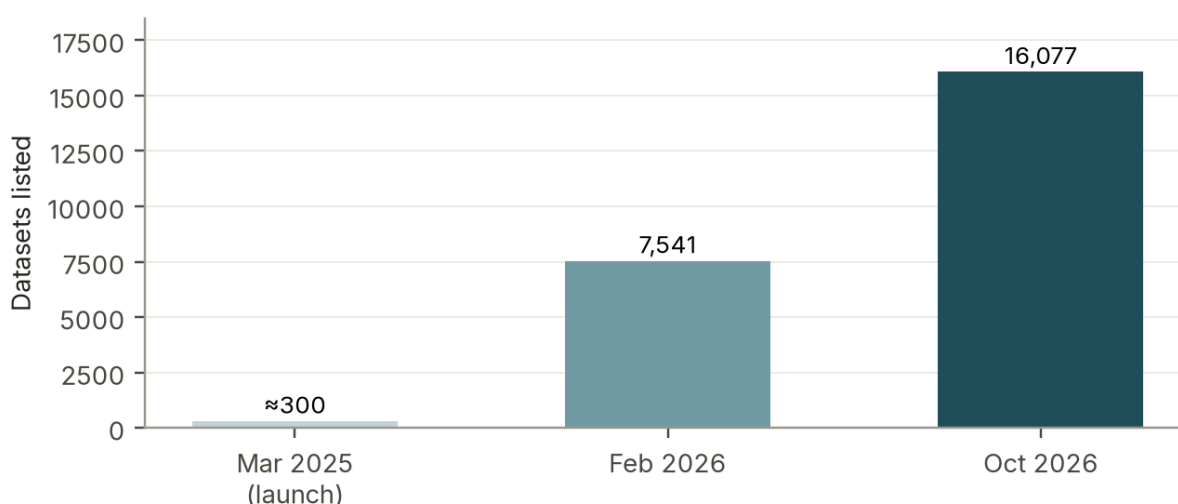


Figure 1: Datasets listed on AIKosh

The launch figure is approximate. Sources: IMPRI; PIB; AIKosh.

1. Khaitan & Co, ERGO: India AI Governance Guidelines, 11 November 2025. <https://www.khaitanco.com/sites/default/files/2025-11/ERGO%20-%20India%20AI%20Governance%20Guidelines%20-%2011%20November%202025.pdf>

2. IMPRI, “AIKosh portal under IndiaAI Mission,” 2025. <https://www.impriindia.com/insights/policy-update/aikosh-portal-under-indiaai-mission/>

3. Press Information Bureau, IndiaAI background note, February 2026. <https://static.pib.gov.in/WriteReadData/specificdocs/documents/2026/feb/doc2026210784401.pdf>

4. AIKosh, IndiaAI datasets platform, accessed October 2026. <https://aikosh.indiaai.gov.in/home/>

Counting datasets says little about their usefulness. Independent reviewers have noted that there is “no data showing the usefulness of this content or whether it was actually helpful in building projects.”⁵ More fundamentally, the most valuable public data — identifiable health records, welfare transactions, land and tax data — cannot be published on a portal at all. It needs a different route: controlled access, under supervision, for defined purposes.

5. IMPRI, “AIKosh portal under IndiaAI Mission,” 2025. <https://www.impriindia.com/insights/policy-update/aikosh-portal-under-indiaai-mission/>

SECTION 2

The legal gap

What the DPDP framework says

The Digital Personal Data Protection Act, 2023 and its Rules, notified on 13 November 2025,⁶ are India's main data protection framework. Most substantive provisions take effect on 13 May 2027.⁷

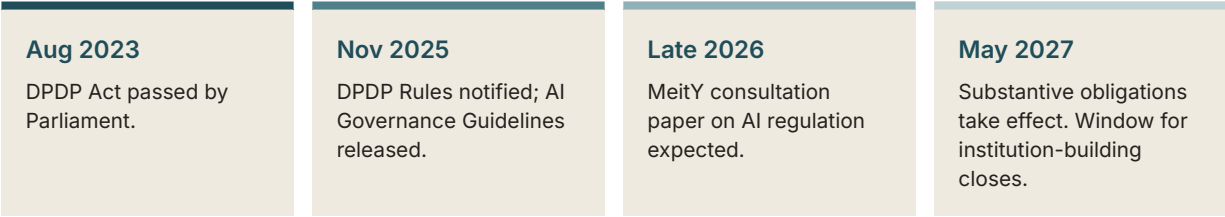


Figure 2: The data governance timeline

The period before May 2027 is when access institutions and guidance on training can be designed without disrupting established practice.

Section 17(2)(b) exempts processing “necessary for research, archiving or statistical purposes if the personal data is not to be used to take any decision specific to a Data Principal,” subject to prescribed standards.⁸ Rule 16 and the Second Schedule set those standards: lawful processing, data minimisation, accuracy, limited retention and reasonable security safeguards.⁹

Three problems

The law never mentions AI

Neither the Act nor the Rules uses the words “artificial intelligence” or “machine learning.”¹⁰ Whether training a model counts as “research” is left to interpretation by each data holder and, eventually, the Data Protection Board. That uncertainty cuts both ways: cautious holders refuse legitimate requests, while others proceed on generous readings that may not survive challenge.

Research becomes deployment

The exemption depends on data not being used to take decisions about individuals. But a model trained for research is commercially and administratively valuable precisely because it can later be applied to individuals: to score credit, flag patients or target welfare. Lawyers already advise that the exemption may be lost once outputs are used for individual decisions.¹¹ The law does not say who checks, or when.

6. DSCI, Digital Personal Data Protection Rules, 2025. <https://www.dsci.in/files/content/documents/2025/Digital-Personal-Data-Protection-Rules-2025.pdf>
7. Legal500, “How does the DPDP Act regulate AI training and automated decision-making in India?”. <https://www.legal500.com/intelligence/india/privacy/how-does-the-dpdp-act-regulate-ai-training-and-automated-decision-making-in-india>
8. DPDP Act 2023, Section 17. <https://dpdpacts.com/chapter-4/section-17/>
9. Lakshmikumar & Sridharan, DPDP Rules: research exemption. <https://dpdp.lkslaw.com/rules/research-exemption>
10. dpdprules.org, frequently asked questions. <https://dpdprules.org/questions>
11. Legal500, “AI training data under India’s DPDP regime: compliance challenges and strategies”. <https://www.legal500.com/intelligence/india/technology/ai-training-data-under-india-s-dpdp-regime-compliance-challenges-and-strategies>

Nobody grants or supervises access

Each data holder, whether a ministry, a state department or a public hospital network, decides alone whether to share data for AI development. There is no permit system, no public register of who accessed what, no common technical standard and no consistent treatment of similar requests.

SECTION 3

Models that work

DEPA for Training (India)

The Data Empowerment and Protection Architecture (DEPA) for Training, developed with iSPIRT and launched in 2023, is a technical and legal design for training AI on sensitive data without copying it.¹² It works in four steps:

- **Registration:** data providers register datasets and their permitted uses.
- **Electronic contracts:** data providers, model developers and other parties sign digital contracts setting permitted uses and privacy budgets — limits on how much can be learned about any individual.
- **Confidential clean rooms:** training runs inside hardware-isolated environments that can prove what code ran on the data. The model developer receives the trained model, not the raw data.
- **Audit:** every training run is logged and can be checked against the contract.¹³

The core idea is that the model comes to the data, not the other way round. This removes the largest privacy risk of AI development, the creation of copies of sensitive data outside the holder's control.

The European Health Data Space

The EU's 2025 regulation on the European Health Data Space (EHDS) focuses on institutions. Each member state must designate a health data access body to review applications and issue data permits. Data is accessed only in pseudonymised or anonymised form inside secure processing environments; personal data cannot be downloaded. Certain uses, including advertising and decisions detrimental to individuals, are prohibited.¹⁴ Training and testing AI systems are explicitly permitted purposes, and individuals have a simple, reversible right to opt out of secondary use.¹⁵

Findata (Finland)

Finland created a working example years before the EHDS. Findata, established in 2019 under the Act on the Secondary Use of Health and Social Data (552/2019), grants permits for secondary use of health and social data held by multiple public bodies. It compiles and pseudonymises the data and provides it inside Kapseli, a secure processing environment. It operates alongside the National Institute for Health and Welfare but separately from the Institute's other activities.^{16,17} That separation between the permit authority and the data users is the institutional feature India most lacks.

12. Digital Impact Alliance, AI x DPI case study: DEPA, May 2026. <https://dial.global/wp-content/uploads/2026/05/AI-x-DPI-Case-study-DEPA.pdf>

13. iSPIRT, DEPA Training technical updates. <https://pn.ispirt.in/depa-training-tech-updates/>

14. EUR-Lex, Summary of the European Health Data Space Regulation. <https://eur-lex.europa.eu/EN/legal-content/summary/european-health-data-space-ehds.html>

15. European Commission, European Health Data Space Regulation. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en

16. Findata, About Findata. <https://findata.fi/en/about-findata/>

17. Findata, Legislation. <https://findata.fi/en/services-and-instructions/legislation/>

Feature	DEPA for Training	EHDS	Findata	India today
Legal basis	Contracts within existing law	EU regulation	National statute	DPDP research exemption, untested for AI
Who grants access	Data provider via e-contract	Designated access body	Independent permit authority	Each data holder alone
Where computation happens	Confidential clean room	Secure processing environment	Kapseli secure environment	Varies; often copies shared
AI training	Core purpose	Explicitly permitted	Permitted within purposes	Unclear
Individual opt-out	Through consent design	Yes	Set by national law	Not defined for research

Table 1. Comparison of data access models.

DEPA supplies the technology. The European models supply the institution. India needs both.

SECTION 4

The critics are partly right

Evidence from Jharkhand

Any case for using public data must face India's own evidence on exclusion. The most rigorous comes from a randomised evaluation of Aadhaar-based biometric authentication in Jharkhand's public distribution system, which supplies subsidised food grain to 15 million beneficiaries. Authentication alone did not reduce leakage. The study's authors report that the zero average impact “masked a 10% reduction in benefits for the 23% of beneficiaries who had not linked an Aadhaar card to benefit rolls, with 2.8% receiving no benefits at all,” and that beneficiaries' transaction costs rose by an average of 17%.¹⁸ A different measure, reconciliation of transaction records, did sharply reduce leakage.



Figure 3: Effects of biometric authentication in Jharkhand's PDS

Transaction costs are averages across beneficiaries; the benefit reduction applies to the 23% without a linked Aadhaar. Source: Muralidharan, Niehaus and Sukhtankar (2020), via Ideas for India.

A related survey found that about 88% of ration cards cancelled in study districts belonged to genuine households.¹⁹ One of the study's authors summarised the evidence this way: “to an extent, both supporters and critics of Aadhaar for service delivery are correct.”²⁰

The lesson for AI

The greatest risk from using public data for AI is not that training data leaks. It is that a model trained on administrative records is then used to make decisions about people, with errors that fall hardest on those whose records are incomplete — often the poorest, most mobile and least connected. Those same people are least able to contest an automated decision.

18. Sukhtankar, S., “Balancing corruption and exclusion: Incorporating Aadhaar into PDS,” Ideas for India, April 2020, summarising Muralidharan, Niehaus and Sukhtankar (2020). <https://www.ideasforindia.in/topics/poverty-inequality/balancing-corruption-and-exclusion-incorporating-aadhaar-into-pds>

19. The Wire, “Aadhaar, PDS and ration cards in Jharkhand”. <https://m.thewire.in/article/politics/aadhaar-pds-ration-cards-jharkhand-jpal>

20. Indian Express, interview with Karthik Muralidharan. <https://indianexpress.com/article/governance/karthik-muralidharan-to-an-extent-both-supporters-and-critics-of-aadhaar-for-service-delivery-are-correct-6283226/>

This is no longer hypothetical. Andhra Pradesh's proposed Family Benefit Management System would develop “AI-based vulnerability scores to improve targeting of government schemes.”²¹ Better targeting is a legitimate goal, and it may well reduce exclusion if designed carefully. It is also exactly where data governance and welfare exclusion meet.

Other objections that deserve weight

- **Surveillance.** Infrastructure that combines datasets for AI also creates the means for state profiling. Clean rooms limit copying but do not by themselves limit purpose.
- **Consent is often not meaningful.** Individual consent rarely reflects real choice when a welfare benefit depends on it.
- **Value capture.** Public data could end up enriching a small number of private model developers, with little return to the public that generated it.

21. Telangana Today, “Andhra's Mana Mitra WhatsApp governance platform expands to offer 1,126 govt services,” July 2026. <https://telanganatoday.com/andhras-mana-mitra-whatsapp-governance-platform-expands-to-offer-1126-govt-services>

SECTION 5

A tiered model for access

Not all public data carries the same risk, and a single rule for all of it would be either too loose or too tight. A tiered model matches the safeguards to the risk.

Tier	Data	Access route	Permitted uses
1. Open	Non-personal and aggregated data: weather, soil maps, published statistics, public documents	Published on AIKosh or departmental portals under open licence	Any lawful use
2. Controlled	De-identified records with residual re-identification risk: anonymised health encounters, crop insurance claims	Permit from access body; compute-to-data in clean room; logged runs	Research and model training for approved purposes
3. Restricted	Identifiable or highly sensitive data: welfare eligibility, detailed health records	Permit plus ethics review; clean room only; privacy budget enforced	Approved public-interest research; no direct export of models without review
Decision use	Any model applied to individuals	Separate impact assessment and approval, regardless of training tier	Only with group-level error testing, human review and appeal

Table 2. A tiered model for access to public data for AI.

Six proposals

1. Clarify the law on model training

MeitY should issue guidance, or amend the DPDP Rules, stating when model training qualifies under the research exemption and when it does not. The guidance should define training, fine-tuning and evaluation; set out what makes a use “specific to a Data Principal”; and specify that the exemption does not carry over to deployment. The AI Governance Guidelines already recommend clarifying how AI development is treated under the DPDP Act. MeitY's forthcoming consultation paper on AI regulation is a natural vehicle.

2. Create independent data access bodies

Start with health and agriculture, where data is rich, the public benefit is clear and international models exist. The body granting access should be separate from the departments that use the data or deploy models, as Findata is separate from the institute it sits beside. It should handle applications, issue permits, operate or certify secure environments and enforce conditions. States could host their own bodies for state-held data, or join a national one.

3. Make compute-to-data the default

For Tier 2 and Tier 3 data, use should take place only inside clean-room environments such as those designed under DEPA for Training, with logged, auditable runs and enforced privacy budgets. Copying raw data out should be the rare, justified exception. This also links to compute policy: public compute capacity under the IndiaAI Mission could host certified clean rooms.

4. Draw a bright line between training and decisions

A model trained under the research exemption should not be used for individual eligibility, scoring or enforcement decisions without:

- a separate, published impact assessment;
- testing for error rates among vulnerable groups, including those with incomplete records;
- human review of adverse decisions;
- a fast, accessible appeal route, including offline options.

Jharkhand shows why fallback routes are not optional. Where an automated system cannot verify someone, the default should be to continue the benefit while the case is resolved, not to stop it.

5. Publish a public register

List every access permit, its purpose, the user, the data tier and the results, including models produced. The EU requires such transparency for data permits. A register lets researchers, journalists and Parliament see how public data is being used, and makes misuse easier to detect.

6. Share the benefits

Models trained substantially on public data should carry public-interest licence terms. Options include free or low-cost use by government, universities and civil society; open release of weights for models trained mainly on public data; or a royalty paid into a fund for public data infrastructure. The aim is that the public that generated the data shares in what is built from it.

Proposal	Owner	Instrument	Timing
Clarify law on training	MeitY	Guidance or Rules amendment	Before May 2027
Independent access bodies	MoHFW, Agriculture, states	Executive order; later statute	Pilot within 12 months
Compute-to-data default	Access bodies, IndiaAI	Permit conditions; certified clean rooms	With first permits
Training–decision line	MeitY, line departments	Guidance; scheme guidelines	Before any AI-based targeting goes live
Public register	Access bodies	Online register	With first permits
Benefit sharing	IndiaAI, access bodies	Standard licence terms	Within 12 months

Table 3. Implementation summary.

SECTION 7

Open questions

This report does not resolve several hard questions, each of which needs consultation with privacy advocates, welfare researchers and state governments, not only technologists:

- **Opt-out for welfare data.** Is individual opt-out workable where the people most at risk may be least able to use it? Would a collective safeguard, such as mandatory group-level testing, protect them better?
- **Funding independence.** How can access bodies be funded without depending on fees from the data users they supervise?
- **Federal allocation.** Should states or the Centre hold access authority for data that states collect under central schemes?
- **Private data.** Should private holders of data generated through public infrastructure, such as payment apps, be brought into the same framework?

SECTION 8

Conclusion

India's choice is not between locking public data away and opening it up. It is between unmanaged access, where each department decides alone and errors are found after harm, and governed access, where institutions, technical safeguards and clear limits on use come first. The technology for governed access is being built in India. The institutions are not. Building them before the DPDP Act's substantive provisions take effect in May 2027 is the opportunity.

About this report

India 2035 Research is an independent research initiative on AI as public infrastructure in India. This report draws on legislation, official releases, academic research and press reporting available as of 24 October 2026. It is not legal advice. Corrections and comments are welcome at rahul@india2035.ai.

Sources

1. Khaitan & Co, ERGO: India AI Governance Guidelines, 11 November 2025. <https://www.khaitanco.com/sites/default/files/2025-11/ERGO%20-%20India%20AI%20Governance%20Guidelines%20-%2011%20November%202025.pdf>
2. IMPRI, "AIKosh portal under IndiaAI Mission," 2025. <https://www.impriindia.com/insights/policy-update/aikosh-portal-under-indiaai-mission/>
3. Press Information Bureau, IndiaAI background note, February 2026. <https://static.pib.gov.in/WriteReadData/specificdocs/documents/2026/feb/doc2026210784401.pdf>
4. AIKosh, IndiaAI datasets platform, accessed October 2026. <https://aikosh.indiaai.gov.in/home/>
5. DSCI, Digital Personal Data Protection Rules, 2025. <https://www.dsci.in/files/content/documents/2025/Digital-Personal-Data-Protection-Rules-2025.pdf>
6. Legal500, "How does the DPDP Act regulate AI training and automated decision-making in India?." <https://www.legal500.com/intelligence/india/privacy/how-does-the-dpdp-act-regulate-ai-training-and-automated-decision-making-in-india>
7. DPDP Act 2023, Section 17. <https://dpdpacts.com/chapter-4/section-17/>
8. Lakshmikumaran & Sridharan, DPDP Rules: research exemption. <https://dpdp.lkslaw.com/rules/research-exemption>
9. dpdprules.org, frequently asked questions. <https://dpdprules.org/questions>
10. Legal500, "AI training data under India's DPDP regime: compliance challenges and strategies". <https://www.legal500.com/intelligence/india/technology/ai-training-data-under-india's-dpdp-regime-compliance-challenges-and-strategies>
11. Digital Impact Alliance, AI x DPI case study: DEPA, May 2026. <https://dial.global/wp-content/uploads/2026/05/AI-x-DPI-Case-study-DEPA.pdf>
12. iSPIRT, DEPA Training technical updates. <https://pn.ispirit.in/depa-training-tech-updates/>
13. EUR-Lex, Summary of the European Health Data Space Regulation. <https://eur-lex.europa.eu/EN/legal-content/summary/european-health-data-space-ehds.html>
14. European Commission, European Health Data Space Regulation. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en
15. Findata, About Findata. <https://findata.fi/en/about-findata/>
16. Findata, Legislation. <https://findata.fi/en/services-and-instructions/legislation/>
17. Sukhtankar, S., "Balancing corruption and exclusion: Incorporating Aadhaar into PDS," Ideas for India, April 2020, summarising Muralidharan, Niehaus and Sukhtankar (2020). <https://www.ideasforindia.in/topics/poverty-inequality/balancing-corruption-and-exclusion-incorporating-aadhaar-into-pds>
18. The Wire, "Aadhaar, PDS and ration cards in Jharkhand". <https://m.thewire.in/article/politics/aadhaar-pds-ration-cards-jharkhand-jpal>
19. Indian Express, interview with Karthik Muralidharan. <https://indianexpress.com/article/governance/karthik-muralidharan-to-an-extent-both-supporters-and-critics-of-aadhaar-for-service-delivery-are-correct-6283226/>
20. Telangana Today, "Andhra's Mana Mitra WhatsApp governance platform expands to offer 1,126 govt services," July 2026. <https://telanganatoday.com/andhras-mana-mitra-whatsapp-governance-platform-expands-to-offer-1126-govt-services>