Data Work in India

Canonical variant. Version 1.0, April 2026.

Preface

About this document. This is a domain-specific variant of Split-Domain Cognition, addressed to the Indian data-work industry — from visual annotation through embodied capture. It is an unusual application: the two domains here, the worker's labour relation and the buyer's data market, have already been kept apart, but in a way that makes the arrangement *unauditable* rather than legible. SDC's commitment is to visibility across domains, not merely their separation. This variant works out what visibility-across-domains means when the domains in question are a labour market and a model market, and what a non-collapsed version of the industry would have to look like. The empirical scope ranges from the roughly seventy thousand annotators currently employed in rural and Tier 2/3 India to the embodied-data-capture frontier projected to follow.

Where this sits in the corpus. The canon home is splitdomaincognition.org. The variants index is at /variants/; this variant in particular at /variants/data-work-in-india/v1/. The principle this variant interprets is articulated long-form in *Split-Domain Cognition* and short-form in *A Principle, Not a Pattern*. The protocol by which this variant was derived is at /derivation-protocol-v1/.

Authority and version. Canonical, v1.0. April 2026. The website is the source of record. If this PDF and the website disagree, follow the website.

Use. Openly citable. See Governance for the one-person canon and how variants are admitted.

Author. Prayas Abhinav.

Opening

An industry has been taking shape in rural and Tier 2/3 India over the past decade — currently around 70,000 workers, projected to approach one million by 2028 — that is almost entirely invisible to the publics whose preferences, voices, images, and now bodily movements it annotates, moderates, records, and captures. The workers describe themselves, and are described in their contracts, as annotators, moderators, recorders, operators, stitchers. The buyers describe the same work in a different frame: as training data for models whose market value exceeds the Indian data-work industry's total wage bill by several orders of magnitude. Between the two descriptions sits a deliberate silence. The worker has no language for the buyer's frame; the buyer has no obligation to translate.

This variant asks what Split-Domain Cognition can say about a labour market structured in this way. It is a slightly unusual application of the principle. In most SDC domains, the collapse to be named is a fusion of language-work and judgement-work into a single act. Here the two are already separated — structurally, legally, geographically. What has failed is not the separation itself but the requirement that both domains be visible to the parties who stand inside either one. The Indian data-work industry is a case in which the domains have been kept apart in order to make the arrangement unauditable, rather than in order to make it more legible. SDC's architectural commitment is visibility across domains, not merely their separation. The variant below works out what that commitment means when the two domains are a labour relation and a data market, and what a non-collapsed version of this industry would have to look like.

The language work

The domain's descriptive material is abundant. It is the worker's daily practice and the frames in which that practice is named and contracted. At the level of task, this is: drawing bounding boxes on images of streets, pedestrians, and surgical fields; labelling objects, relations, sentiments, and harms in text; reviewing up to 800 violent or pornographic videos per day for conformity to moderation guidelines; recording one's voice in under-resourced Indian languages for speech-recognition datasets; labelling pairs of model outputs by preference for RLHF pipelines; and — in the newest tier — stitching garments while wearing a head-mounted camera that captures first-person video of one's hand movements.

At the level of contract, the language is the language of wages and piece-rates: ₹10,000–15,000 per month for salaried annotators at the lowest tier, ₹25,000 in the best cases,

₹40,000+ for RLHF preference labelling, piece rates that transfer quality risk to the worker, NDAs that foreclose discussion of the work with anyone outside it. At the level of geography, it is the language of rural and Tier 2/3 Indian districts — Ranchi, Shillong, Vishakhapatnam, Bengaluru factory belts — where the cost of living and the scarcity of alternatives are both the reason the work was brought there and the reason the worker cannot refuse it.

These are the materials the domain attends to. A stakeholder outside the domain, given access to them, could describe the work fairly. The description would not be wrong. It would simply be a description of one domain.

The judgement work

The domain also has a judgement layer, but the judgement sits almost entirely outside the worker's frame. The verdicts being issued are verdicts on the work's economic value — to whom, for what, at what conversion rate between one worker-hour and one unit of training-data throughput. The buyer's decision is: this annotation corpus is adequate to fine-tune a model worth X; this moderator's review-rate is adequate to meet a platform's SLA; this egocentric video set is adequate to train a robotic manipulation policy that will be sold to an apparel manufacturer for Y. The further verdicts are about the labour arrangement itself: whether the worker's consent is meaningful, whether the compensation is a fair share of the downstream value, whether the extraction is permissible.

Neither set of verdicts is issued in the worker's presence. Both sets are issued between the contracting platform, the buyer, and the buyer's investors. The worker is the subject of the judgement but not a party to it.

This is not a case where the judgement work is absent. It is a case where it has been held somewhere else, under a different lexicon, with different stakeholders, and with no requirement that the worker be able to read its outputs. The judgement domain exists and is vigorous. What is missing is any bridge between the two domains through which the worker might see what their labour is being priced against.

The collapse, with examples

The SDC collapse in this domain is not a fusion but a structured opacity. The language and judgement domains operate in separate channels, and the separation is the mechanism by which the arrangement is kept unauditable. Four concrete instances make the structure visible.

The Ranchi moderator. A rural Adivasi woman is hired to moderate content. Her contract is a language-frame contract: wage, hours, NDA, quality thresholds. The judgement-frame — that her labour trains a system whose market capitalisation depends on her throughput, that her trauma has a dollar value to a platform whose scale her district will never be shown — is held in the buyer's frame, in English, in a jurisdiction she will not access. She cannot discuss either frame with anyone, because the NDA forecloses the first and the second was never available to begin with.

The RLHF labeller. A Bangalore graduate labels preferences between model outputs for ₹40,000 a month. She knows what RLHF is. She does not know which model ships, which market is contested, which labour-displacement downstream benefits from her labelling work. She is nearer the buyer's frame than the Ranchi moderator but still inside a partial view of it: the act is named, the economic architecture is not.

The Karya voice-recorder. A worker records thirty seconds of speech in an under-resourced Indian language for an above-minimum wage, with the recording explicitly framed as training data and with a data-ownership claim retained by the worker and the cooperative. Here the two domains have been made visible to the worker by design. Karya is the domain's clearest existing example of what the non-collapsed version of this industry can look like. It is also, for reasons worked out below, the hardest version of the architecture to scale.

The Bengaluru garment worker. The worker is paid to stitch fabric. A head-mounted camera records egocentric video of his hand movements. The recording is sold as training data to a robotics firm whose product will, on the industry's own trajectory, eventually do the work the worker is currently paid to do. The worker is paid against one output and is silently the source of a second. The second output is not named in his contract, not priced against any reference his wage can be compared to, and not disclosed to him as a good whose value has been realised. The collapse is so complete that the second frame is not merely invisible to the worker but ambient — indistinguishable to him from the act of working itself.

In each of these cases the structural move is the same. The work is described in the language of one contract. The value is adjudicated in the frame of another. The asymmetry between the two frames is maintained by NDAs, contractual segmentation,

geographic isolation, language barriers, and the sheer scale of the buyer's architecture. The worker's inability to move between the two frames is not a side-effect; it is the operating principle.

The cost

The cost of this arrangement is borne across several registers. The most immediate is financial. Wages are set against what the language-frame can support (a local labour market with few alternatives) rather than against what the judgement-frame can bear (a global data market whose downstream values are reported in billions). The gap between the two determines the extractive margin, and the gap is maintained by the domains being kept apart.

A second cost is psychological and physical. The content moderator absorbs the platform's harm at scale — up to 800 items per day, in shifts of up to twenty hours, with mental health support available in a minority of firms — because the frame in which her labour is priced has no category for psychic injury. Mental health support would be a translation from the judgement frame (what this labour does to a human over time) into the language frame (what the contract acknowledges). The absence of the translation is not an oversight. It is what keeps the arrangement economically possible at the current wage.

A third cost is political. A workforce approaching a million people, concentrated in rural and Tier 2/3 India, disproportionately Dalit and Adivasi, disproportionately female in the moderation tier, is being formed without the public scrutiny such a workforce would normally attract. The NDA and the geography work together: the NDA forecloses speech, the geography forecloses witness. The result is an industrial-scale labour relation that is, by design, unable to tell its own story.

A fourth cost is civilisational, and it is the one the fifth tier makes most visible. The embodied-capture worker is paid to perform an act of skill that is simultaneously being acquired as training material for systems designed to eventually replace him. The silence about the second extraction is not a background feature of the arrangement; it is the condition of its possibility. If the worker knew, the contract would have to change. If the contract changed, the arrangement would cease to be profitable at current scale. The domain is built on the asymmetry, and the domain will lose its current shape the moment the asymmetry is closed.

The separation — what SDC would require

What would the domain look like if the SDC commitment were taken seriously — not as a rhetorical claim but as an architectural requirement? The five deliverables from `../process.md` (Domain Map, Language Instrument, Rubric, Narration Mechanism, Audit Protocol) can be named in the domain's own terms.

The Domain Map. The full shape of the transaction is laid out: who contracts whom, who pays whom, which outputs flow to which buyers, at what conversion rates. The map is made visible to all parties — the worker, the contracting platform, the downstream buyer, and any auditor. At present the map exists only in the buyer's frame. In a non-collapsed arrangement it would exist in the worker's frame as well.

The Language Instrument. The worker is given an account of their labour in language they can recognise. Not a translation of the buyer's contract into the worker's mother tongue (though that too), but a description that names both outputs of the labour where two outputs exist. For the garment worker, this means naming that the shift produces both garments and egocentric video, with both named on the wage slip. For the content moderator, it means naming that the shift produces both cleaned feeds and training signal for automated moderation systems, and that the labour's psychic cost is recognised in the contract.

The Rubric. The criteria by which the arrangement is judged acceptable are made explicit before the case is decided. Fairwork's five-criterion framework (pay, conditions, contracts, management, representation) is the domain's closest existing instrument; an SDC-aligned rubric would extend it to cover the data-extraction layer specifically — consent for each distinct output, compensation that tracks the downstream value, a right of withdrawal that does not void the worker's underlying labour contract.

The Narration Mechanism. When a verdict is issued — this arrangement is acceptable, this contract is renewable, this worker's data may be sold to this buyer — the verdict is narrated back to the worker in terms that make its reasoning reconstructable. The buyer's frame and the worker's frame are bridged by an explicit account of why the rubric was met. Refusal of the bridge becomes visible as refusal, rather than dissolving into the opacity that is currently the default.

The Audit Protocol. An outside party can verify that the separation has been honoured. The records required for audit exist, and they contain the information needed to reconstruct both the language frame (what the worker was contracted to produce) and

the judgement frame (how the output was priced downstream). Karya's cooperative-ownership architecture is the closest existing proof that this is possible. The harder tiers — moderation, visual annotation, and above all embodied capture — would require contract architectures that do not yet exist.

The separation SDC requires in this domain is not the separation the industry already performs. The industry's separation is weaponised opacity. The SDC separation is illuminated legibility. The difference is whether both domains are inspectable by all parties, or whether one party is confined to a single side of the division.

Philosophical grounding

The deepest philosophical underpinning for this variant is the argument in `worth-is-not-hierarchical.md` (Prayas Abhinav, 11 April 2026). That document shows that hierarchies of worth across persons fail at the level of the common unit: different domains of human excellence do not share a unit in which a weighted sum could be computed. The argument grounds the animal-rights variant by extending the incommensurability claim across species. It also grounds the present variant in a different way.

The Indian data-work industry's implicit claim is that the worker's labour-hour and the buyer's model-capability operate on a shared unit — the rupee, the dollar, the per-unit cost of compute-hour equivalents. The industry asserts that a local labour market's clearing price is the correct price for the labour's contribution to a globally-sold model. The worth-is-not-hierarchical argument allows the response to be given structurally rather than politically. There is no common unit across the two domains against which the equivalence can be computed. The rupee figure is a price, not a measure of contribution. The claim that the worker's wage is adequate because the local labour market supports it is a domain-specific claim made from within one frame, smuggling across itself the authority of a second frame (global data markets) whose unit it cannot access. This is the SDC collapse at the scale of political economy.

A second grounding is Karl Polanyi's argument that labour is a fictitious commodity — that the institution which treats human activity as a market-clearing good is not discovering a natural price but imposing one. Polanyi is not cited in the sdc.md canonical articulation, but this variant is in his lineage. The claim that the wage emerges from a market assumes that the market has access to the full shape of what is being exchanged. When the data-

extraction frame is held outside the worker's reach, the market has no such access. The market is not pricing the work; it is pricing the half of the work the worker is permitted to see.

A third grounding, narrower and more operational, is the Fairwork Foundation's labour-conditions framework. Fairwork does not carry SDC's full architectural weight but demonstrates that the move is instrumentable at scale: nearly a decade of Fairwork ratings across gig-economy and cloudwork platforms shows that explicit criteria, applied in public, produce measurable shifts in firm behaviour. SDC's architectural commitment is in the same family; Fairwork is one of its existing empirical expressions.

Relationship to Koher output

This variant does not yet correspond to a Koher tool. Tools in the Koher catalogue so far address the collapse as it appears in pedagogy (studioMeetingCcompanion), ethics (animalRightsLens), and cognitive legibility (Coherence Diagnostic, Fragment Mapper, Play Shape Diagnostic). The data-work domain is the first Koher variant in which the collapse is a labour relation rather than a cognitive act, and in which any resulting instrument would have to live in the contractual and auditing layer rather than in a Stage 1 / Stage 2 / Stage 3 pipeline.

The nearest adjacent work is the ongoing interest in Karya, which is a living demonstration of the non-collapsed arrangement in one tier. The notes file `../notes/embodied-data-capture-domain.md` (20 April 2026) explores a narrower version of this variant focused on the fifth tier alone. The monthly concept note at `~/Dropbox/desktop_monthly/.Apr_2026/data/concept-note-problematic-emergence-india-data-industry.md` is the structural source this document is drafted against.

Three possible Koher landings are available if this variant is taken further:

- A public-facing instrument modelled on Fairwork — a data-work-specific rubric that rates firms on whether the language and judgement domains have been made mutually visible to their workers. This would be a partnership rather than a solo Koher build.

- An artwork sibling to *The Murmur Engine* — a public installation that renders the data-extraction layer visible at the site of the labour, in the worker's language, in a form the worker can re-enter. The FICA submission architecture has the right shape; the execution would be in a factory context rather than a gallery.
- A writing thread that makes the structural argument available to the Indian policy and labour-law conversation, which remains framed largely in gig-economy vocabulary and has not yet absorbed the data-extraction layer as a distinct regulatory object.

None of these is committed to. The variant is written here so that the argument is available to the practice when the question of which landing to take up becomes live.

Closing

The Indian data-work industry is one of the clearest existing instances of a labour market built on the SDC collapse in its dual form: not as the fusion of language and judgement into one channel, but as the asymmetric opacity of the two channels to the party whose labour makes both possible. The variant exists to make that structure sayable in the domain's own terms, so that the next argument — regulatory, cooperative, artistic, or curricular — can be made without first having to invent the vocabulary. SDC does not prescribe which intervention to make. It names the shape any honest intervention would have to honour: both domains visible to both parties, both outputs priced against their downstream value, both kinds of labour — the contracted and the ambient — named where the contract is signed.

The shape is not yet present in the Indian data industry at scale. Whether it can be made present is the question the next phase of this work, if it is taken up, will have to answer.

Draft 1 — 20 April 2026. Variant status: drafted but not yet reviewed. To be tested against the four validation tests in `README.md` *: recognition, specificity, cost, and stake. The stake test is the one most open at present: Prayas's position at Anant, his Hype Studies teaching, his existing interest in Karya, and his Ahmedabad-to-Bengaluru geography are the candidate loci of stake. The recognition test awaits a reading by a practitioner in the data-work industry.*