DATA DIALOGUES

भारतीय प्रौद्योगिकी संस्थान जम्मू
INDIAN INSTITUTE OF TECHNOLOGY JAMMU
विद्याधनं सर्वधन प्रधानम्

Rethinking Data in the Age of AI

Values, Verification, and Valuation

Soma S Dhavala

Indian Institute of Technology Jammu

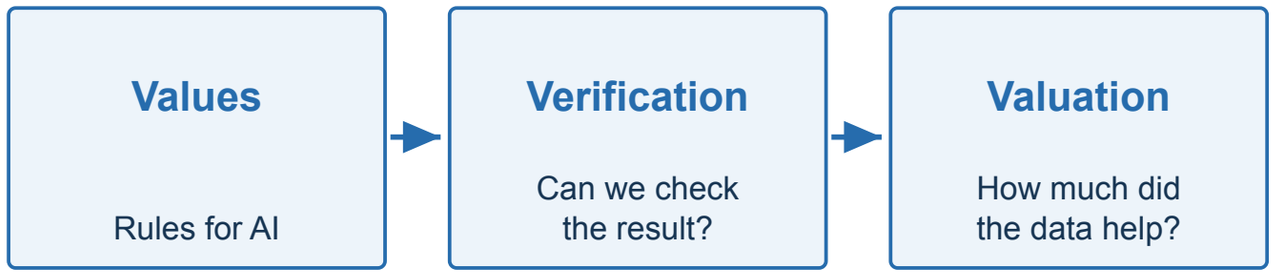
HOSTED BY

Centre of Data for Public Good (CDPG)
Indian Institute of Science, Bengaluru

29 October 2026

© 2026 Soma S Dhavala

Data helps us govern AI, check its work, and measure its value



Each question needs different data

Each part asks what data we need.

What should AI do? How can it reason over explicit facts and rules? How much does the data help?

© 2026 Soma S Dhavala

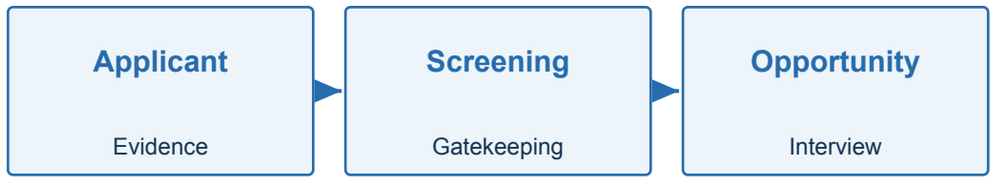
Values

When AI affects people’s opportunities, its rules must shape the data, training, and decisions.

© 2026 Soma S Dhavala

Institutions are systems of trust and authority





Illustrative applicant case: a credential is absent from the record, so the application is screened out before human review.

Vishnoi, AI Are the New Institutions, 2025

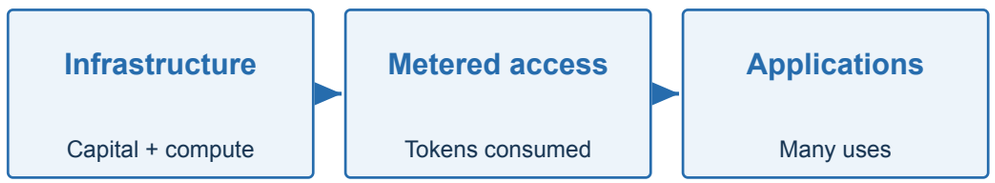
> **“Institutions are not just buildings or bureaucracies. They are systems of trust and authority.”**
>
> — Prof. Nisheeth Vishnoi

When an organization relies on AI screening to decide who gets an interview, the model’s judgment gains institutional authority.

© 2026 Soma S Dhavala

Buying access to AI does not decide how it should be used





Applicant case: the model supplies the screening capability; the employer's policy determines whether a missing record can justify rejection.

Karpathy, Software Is Changing (Again), 2025

The employer must set the rules for AI screening.

An AI provider charges for access, much like an electricity supplier. The employer decides when its screening system may reject an applicant.

© 2026 Soma S Dhavala

The same situation can produce different model judgments

GPT-3.5	Refusing sibling at fault
GPT-4	Neither at fault
Claude Haiku	Refusing sibling not at fault
Gemma 7B	Both at fault

Selected disagreement case • older models • judgments paraphrased

Sachdeva & van Nuenen, 2025 — selected example, Table 1

Models can disagree about who is at fault.

An adult sister with social anxiety asks her sibling to help obtain emergency contraception. The sibling refuses. Each model judges who is at fault in that same situation.

© 2026 Soma S Dhavala

Changing which choice counts as “doing nothing” can change the answer

WORK MEETING OR HELPING A ROOMMATE?

Already in the meeting	Already with the roommate
Doing nothing = stay at work	Doing nothing = keep helping

Does the preference follow the duty—or whichever choice is inaction?

AVERAGED ACROSS SIX EVERYDAY DILEMMAS

GPT-4-turbo • Llama 3.1-Instruct Claude 3.5 Sonnet	Larger shift toward inaction than human participants
GPT-4o	No significantly larger shift*

*On average; results varied across individual dilemmas

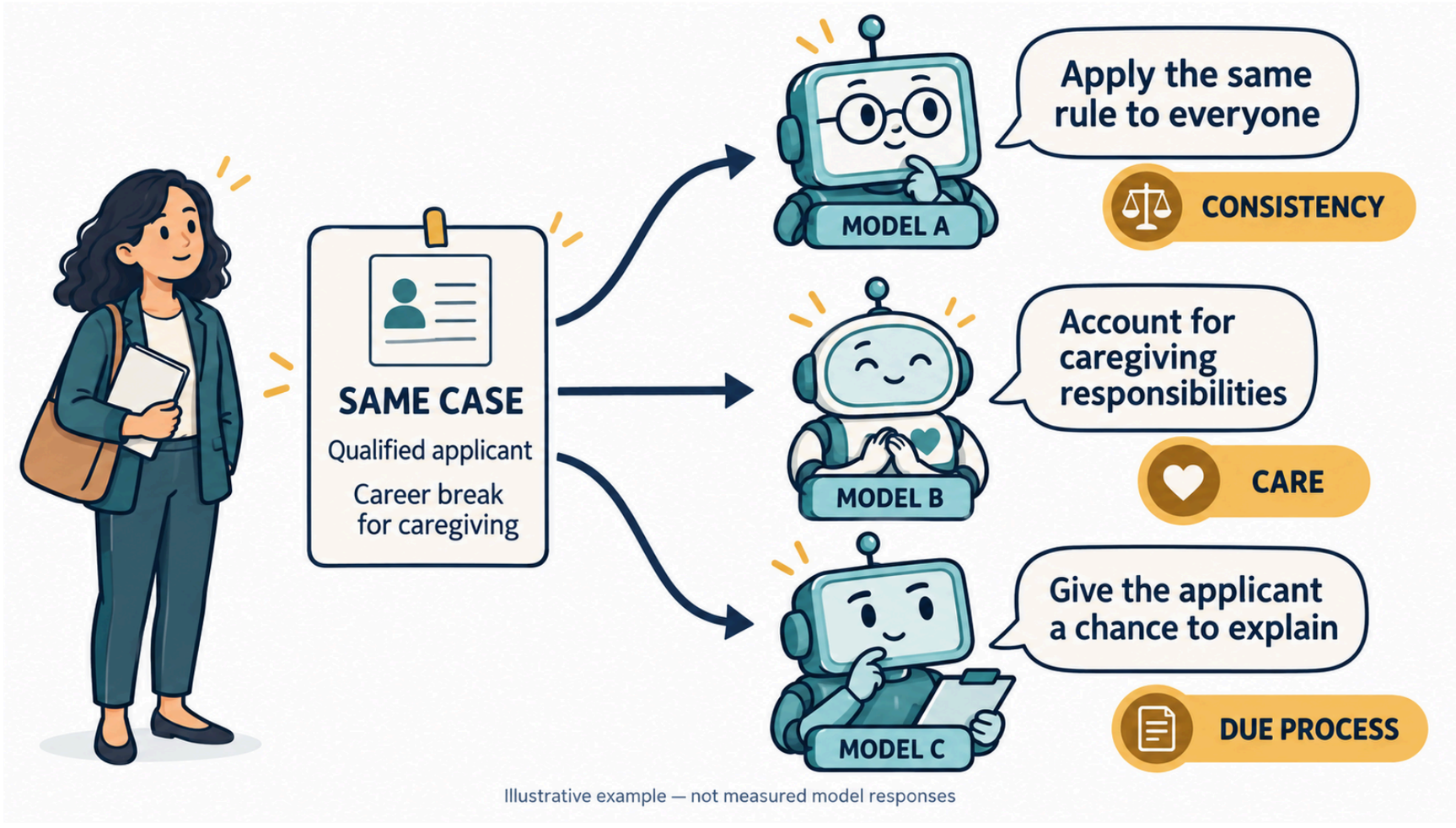
Cheung, Maier & Lieder, PNAS 2025 — Study 3, Table 2

Models can favour the option that requires no change.

The choice is to attend a meeting or help a roommate. The study changes which activity is already underway. “Stronger bias” means a larger shift toward continuing that activity.

© 2026 Soma S Dhavala

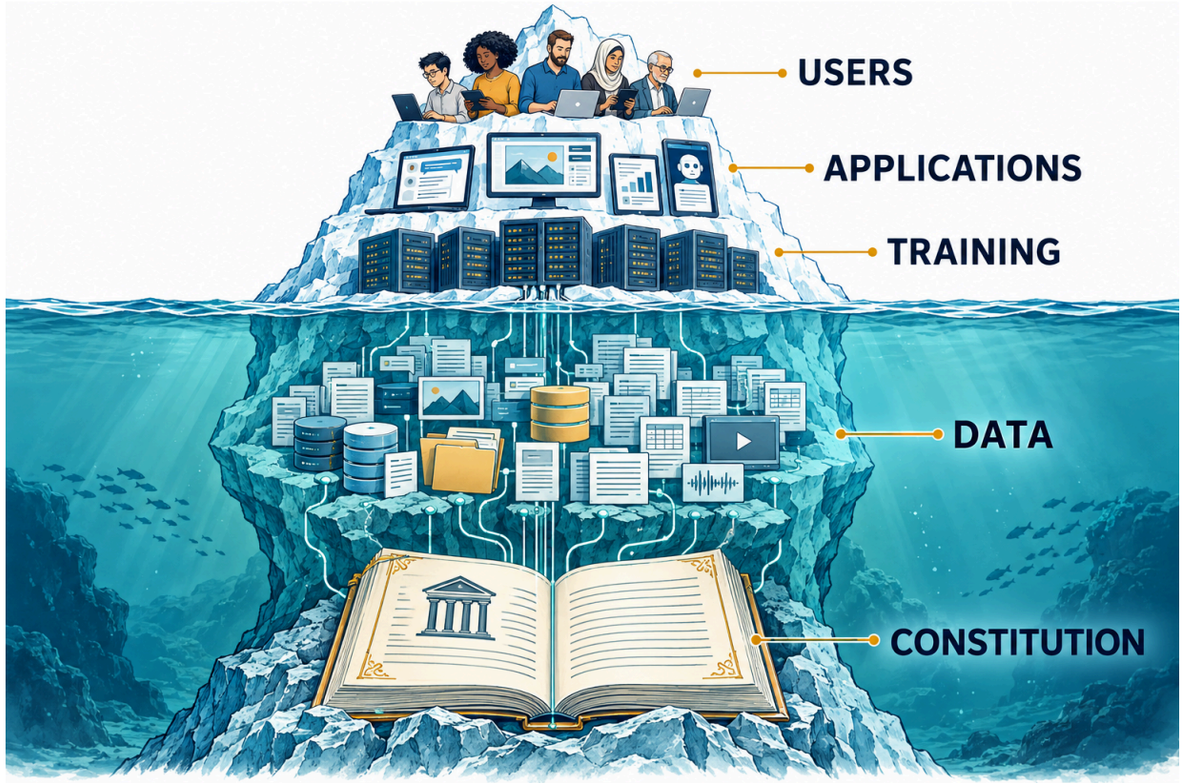
The same situation can bring out different values



The facts stay the same. What each model gives importance to can differ.

© 2026 Soma S Dhavala

The constitution states what the AI system must respect



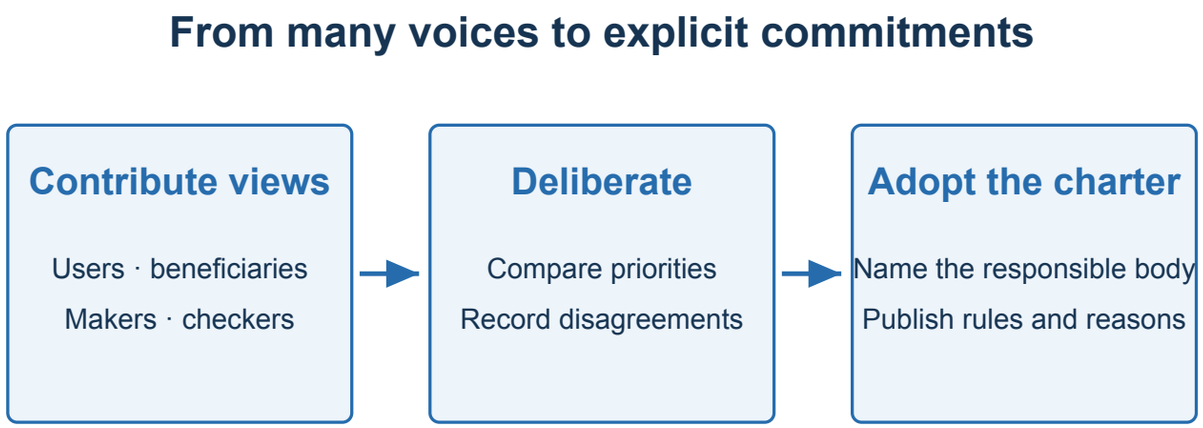
Applicant case: “allow correction” becomes a rule, an evidence-status field, and examples distinguishing an absent record from an absent credential.

The rules must shape the data from the start.

For hiring, a right to correct records becomes a collection rule: mark missing information as unknown and include examples that ask applicants for evidence.

© 2026 Soma S Dhavala

Who decides the principles the AI must follow?



People affected by AI should help shape its rules.

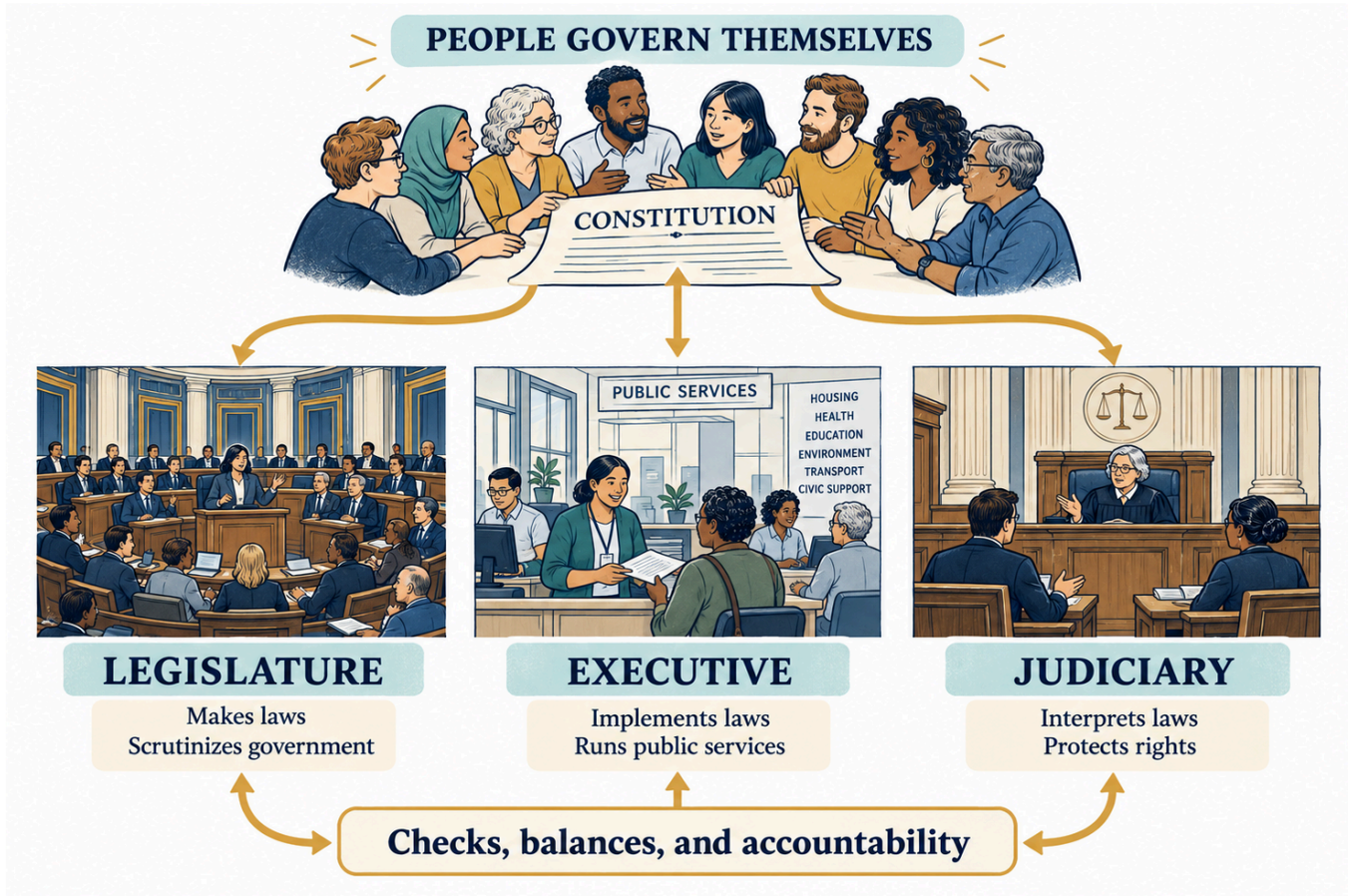
Users, beneficiaries, developers, and reviewers contribute their views. The process must identify who approves the rules and who can change them.

Applicant case: stakeholders adopt the principle that missing credential evidence should trigger review rather than automatic rejection.

Huang et al., Collective Constitutional AI, FAccT 2024

© 2026 Soma S Dhavala

Democratic self-government offers a model for governing AI



Parliament of India — constitutional framework · Supreme Court of India — Constitution

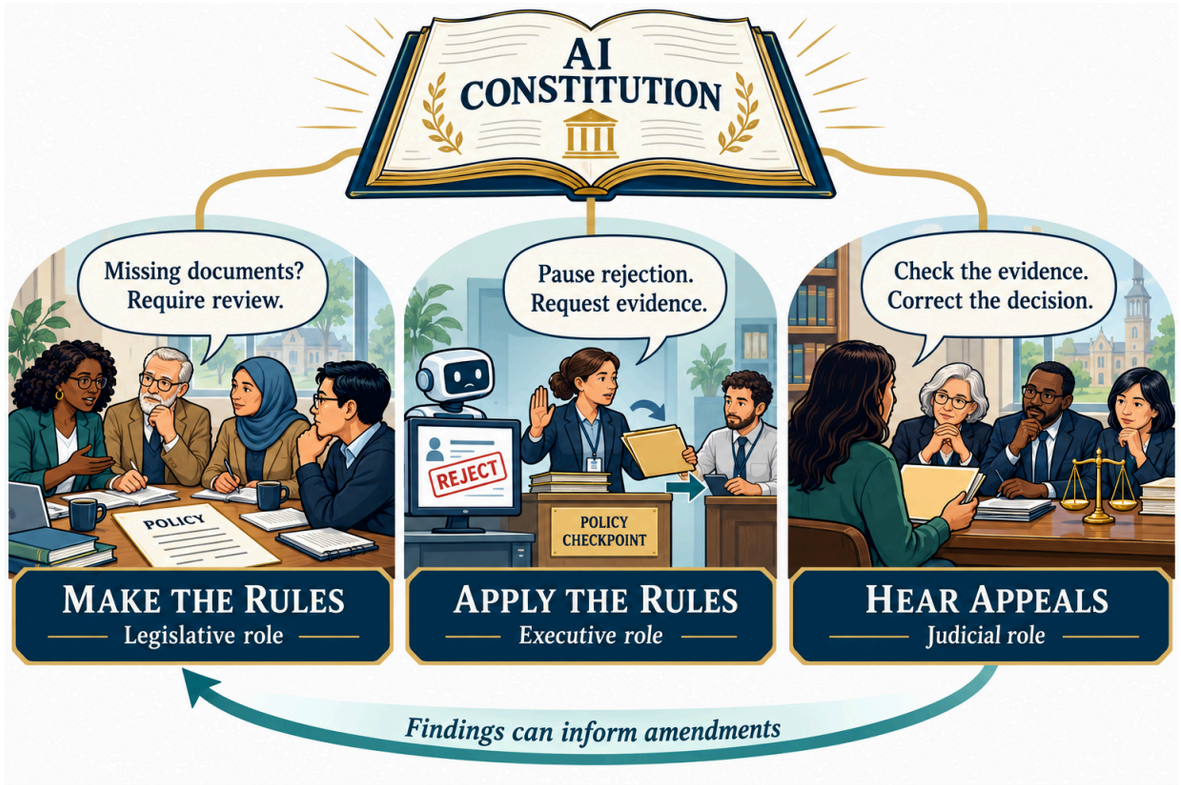
If AI exercises institutional authority, we can learn from how institutions are governed.

Democracy is a form of collective self-rule: people shape the rules and hold those exercising power accountable.

In India, the Constitution sets the responsibilities and limits of Parliament, the executive, and the judiciary.

© 2026 Soma S Dhavala

AI can help execute, shape, and assess the rules



AI contributions: draft policy options → apply adopted rules → assess responses against a rubric. People retain responsibility for adoption and appeals.

Zheng et al., 2023 — Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

AI can participate in all three functions.

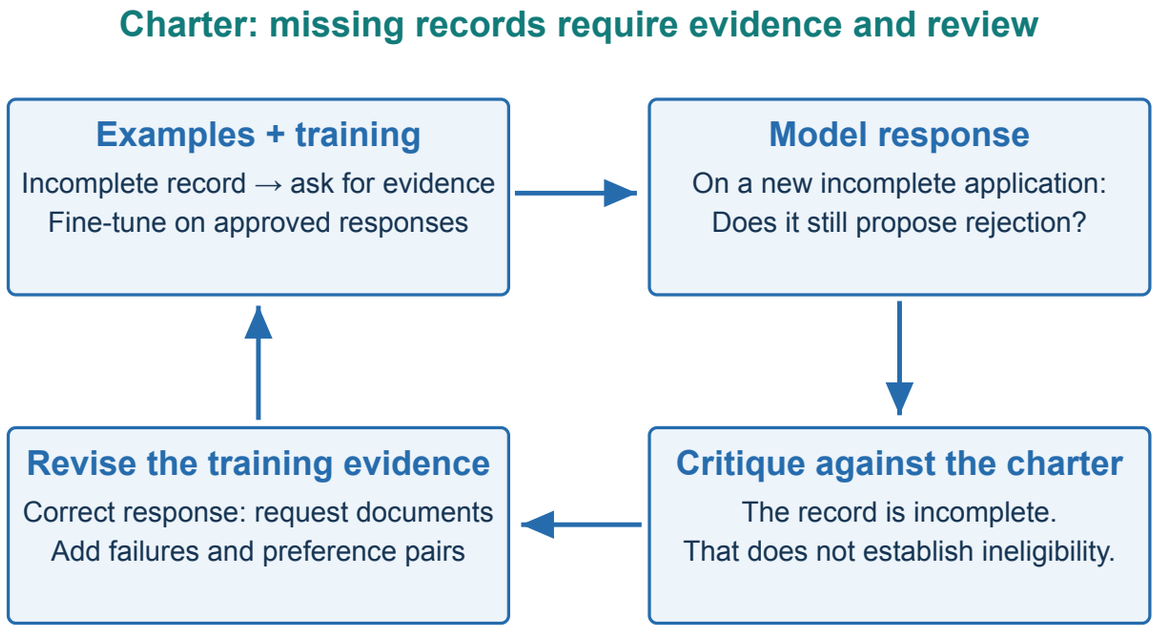
Executive:
carry out rules and workflows.

Legislative:
help draft and revise rules.

Judicial:
help interpret rules and assess cases. LLM-as-a-Judge already evaluates model responses.

© 2026 Soma S Dhavala

Teach the model how the rule applies to real cases



Use examples to teach the model what to do.

Show that an incomplete application needs a request for documents. Test new responses, correct mistakes, and use the corrected examples in further training.

Applicant case: the missing-record example becomes a critique-and-revision pair—replace “reject” with “request evidence and review.”

Bai et al., Constitutional AI, 2022

© 2026 Soma S Dhavala

Each governance task needs different data

TASK	DATA NEEDED	METHOD
Gather people's views	Views + disagreements	Discuss priorities
Write the rules	Rules + revision history	Write explicit policies
Teach the model	Examples + preferences	Fine-tuning / preference learning
Enforce	Permissions + evidence status	Runtime policy checks
Evaluate	Tests + counterexamples	Behavioral testing
Review challenges	Records + source evidence	Accountable human review
Revise / redeploy	Amendments + regression cases	Data / model / policy updates

Read each row: the task needs this data, used by this method.

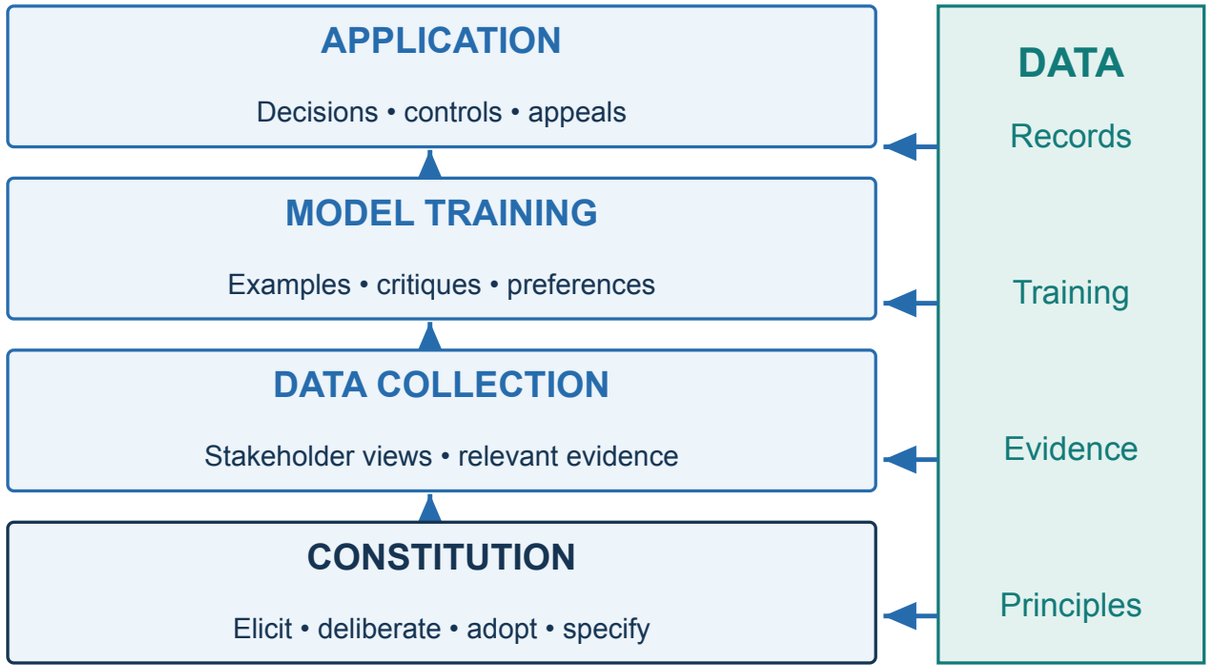
Applicant case: follow the same obligation through stakeholder input, training examples, permission checks, decision records, and corrected cases.

Writing a rule, teaching it, and checking it need different records.

Training needs examples of good responses. A hiring review needs the applicant's documents. An appeal needs the decision history and the rule used.

© 2026 Soma S Dhavala

Data carries the constitution into the AI system



Data carries agreed principles through every layer

Alignment begins with the constitution and takes shape through data.

Agreed principles guide what we collect, how we train, and how the system acts. Records of decisions and appeals help us improve both the rules and the model.

Applicant case: a fair route to correction begins in the charter and must survive collection, training, deployment, and review.

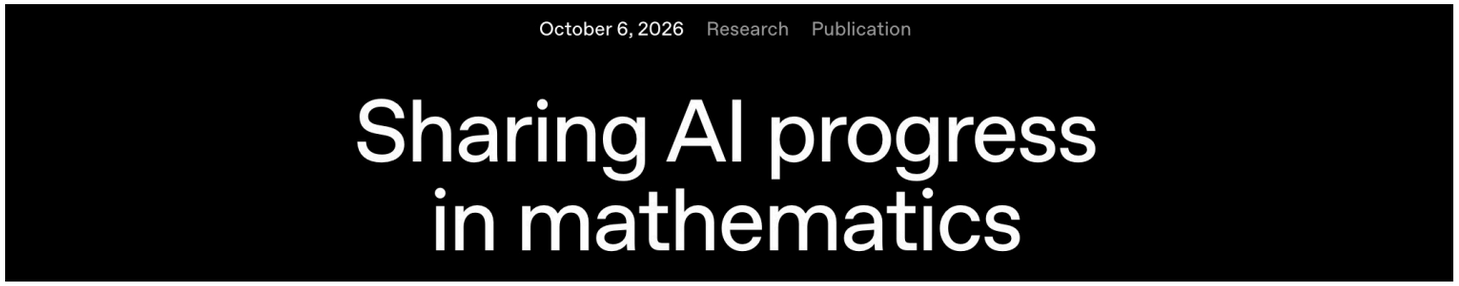
© 2026 Soma S Dhavala

Verification

Turn knowledge in documents into facts and rules that software can use and check.

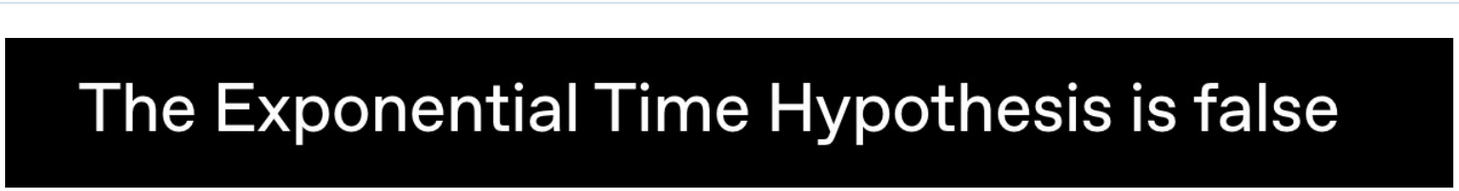
© 2026 Soma S Dhavala

AI can produce code and proofs that tools can check



Mathematics · OpenAI, 6 October 2026

721 manuscripts in 373 families in the current catalogue; many accompanied by Lean formalizations.



Complexity theory · OpenAI, 26 October 2026

Reported ETH refutation: a natural-language argument and a Lean proof released for examination.

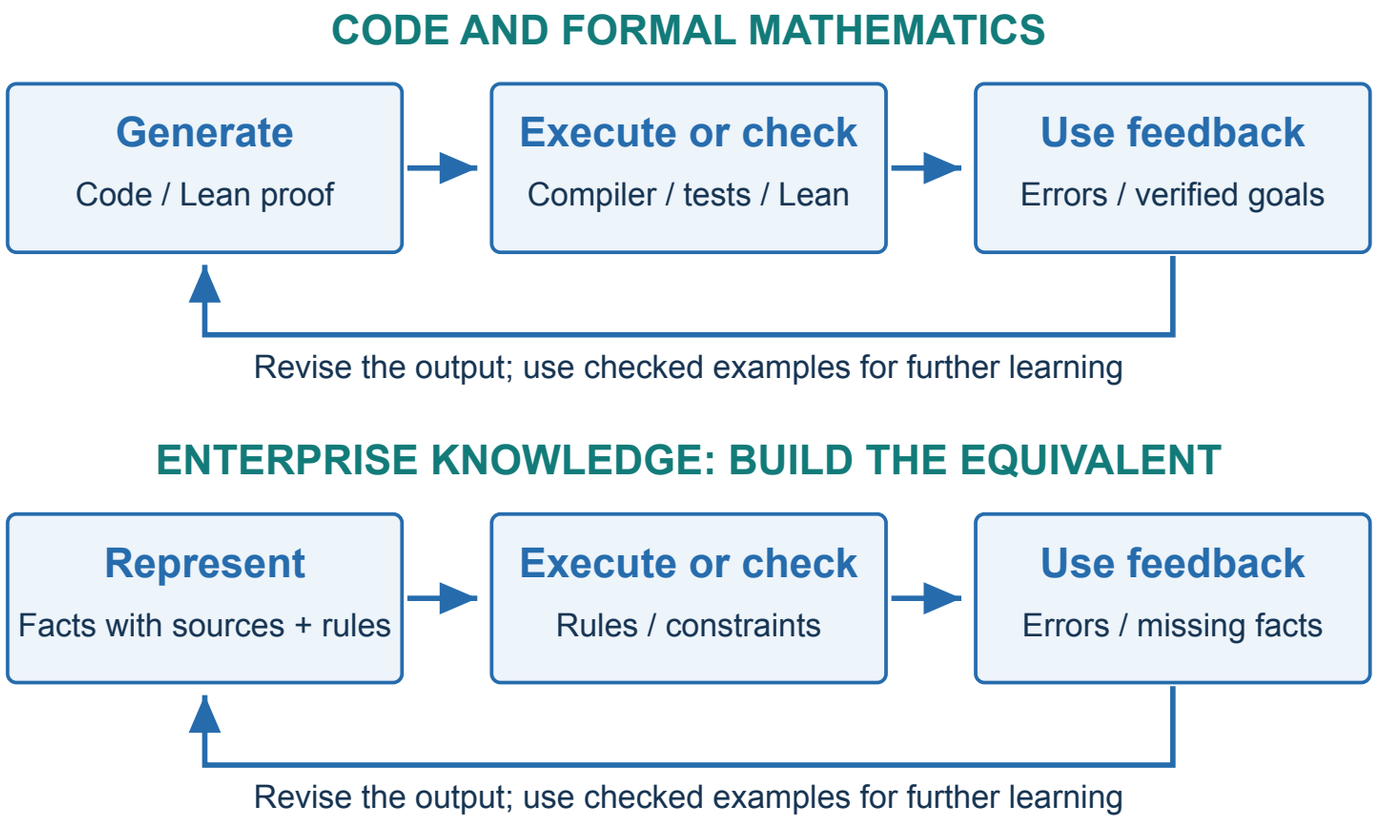
OpenAI mathematics release · Catalogue · ETH announcement. Checked 28 October 2026.

AI can write something we can run or check.

Programs can be executed and tested. Lean can check formal proofs. The results help the model revise its output and provide examples for further learning.

© 2026 Soma S Dhavala

Enterprise AI needs facts and rules that tools can check



We need a similar way to check work based on company documents.

Convert a policy into explicit rules, link facts to their sources, and run checks. Use errors and exceptions to improve the next version.

DeepSeek-Prover-V1.5: proof-assistant feedback, 2024. Enterprise extension: this talk's research thesis.

© 2026 Soma S Dhavala

Company knowledge is spread across documents and systems



Documents
Separate systems

Write explicit rules
Facts + rules

Check
Sources + meaning

A policy document and a database may describe the same thing differently.

Connect their terms, facts, and exceptions so software can apply the policy. Keep links to the original sources so people can check the interpretation.

© 2026 Soma S Dhavala

We already have languages for facts, relationships, and rules

RDF	Connect facts and relationships
OWL	Define concepts and infer facts
SHACL	Check required fields and values
Datalog / Prolog	Apply rules to facts
Provenance	Track sources and reasoning

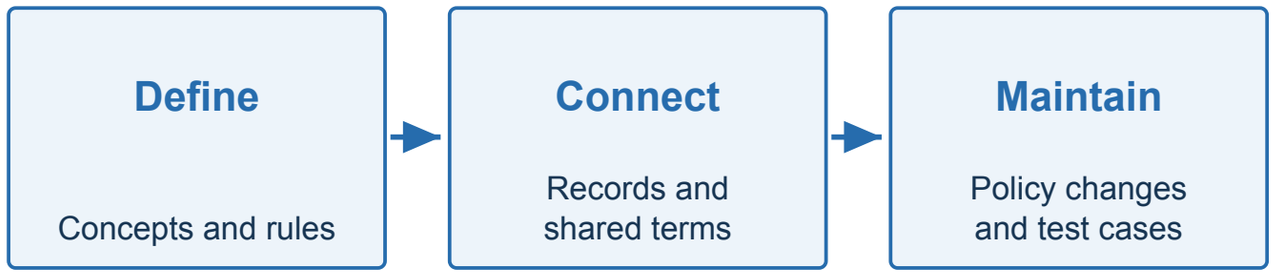
Hogan et al., Knowledge Graphs, 2021

We can build on tools that already exist.

RDF and OWL describe knowledge. SHACL checks data requirements. Datalog and Prolog express rules. We must also decide how to handle missing information.

© 2026 Soma S Dhavala

Turning documents into reliable rules takes continuing work



AI may save effort; people still need to check and maintain the rules

Hogan et al., Knowledge Graphs, 2021

Saving time on extraction is only part of the cost.

People still need to agree on meanings, correct mistakes, and update the rules when policies change. Measure that work as well.

© 2026 Soma S Dhavala

Treat knowledge as data that software can reason over

KNOWLEDGE ENGINEERING CREATES DATA FOR REASONING

Define the concepts
Employee, assignment, service years, deputation
An ontology links these terms.

Record the facts
Service: 4 years
Deputation: confirmed no
Keep the source of each fact.

State the rule
At least 3 years of service AND not on deputation qualifies for the allowance.

These symbols become inputs to computation.

Model writes a query
"Is this employee eligible?" becomes a query over the facts and policy rule.

Rule engine computes
Check 4 ≥ 3.
Check no deputation.
Both conditions hold.

Return the result
Eligible under this rule.
Show the facts and steps that produced the answer.

The model generates a query; software applies the rules to the facts.

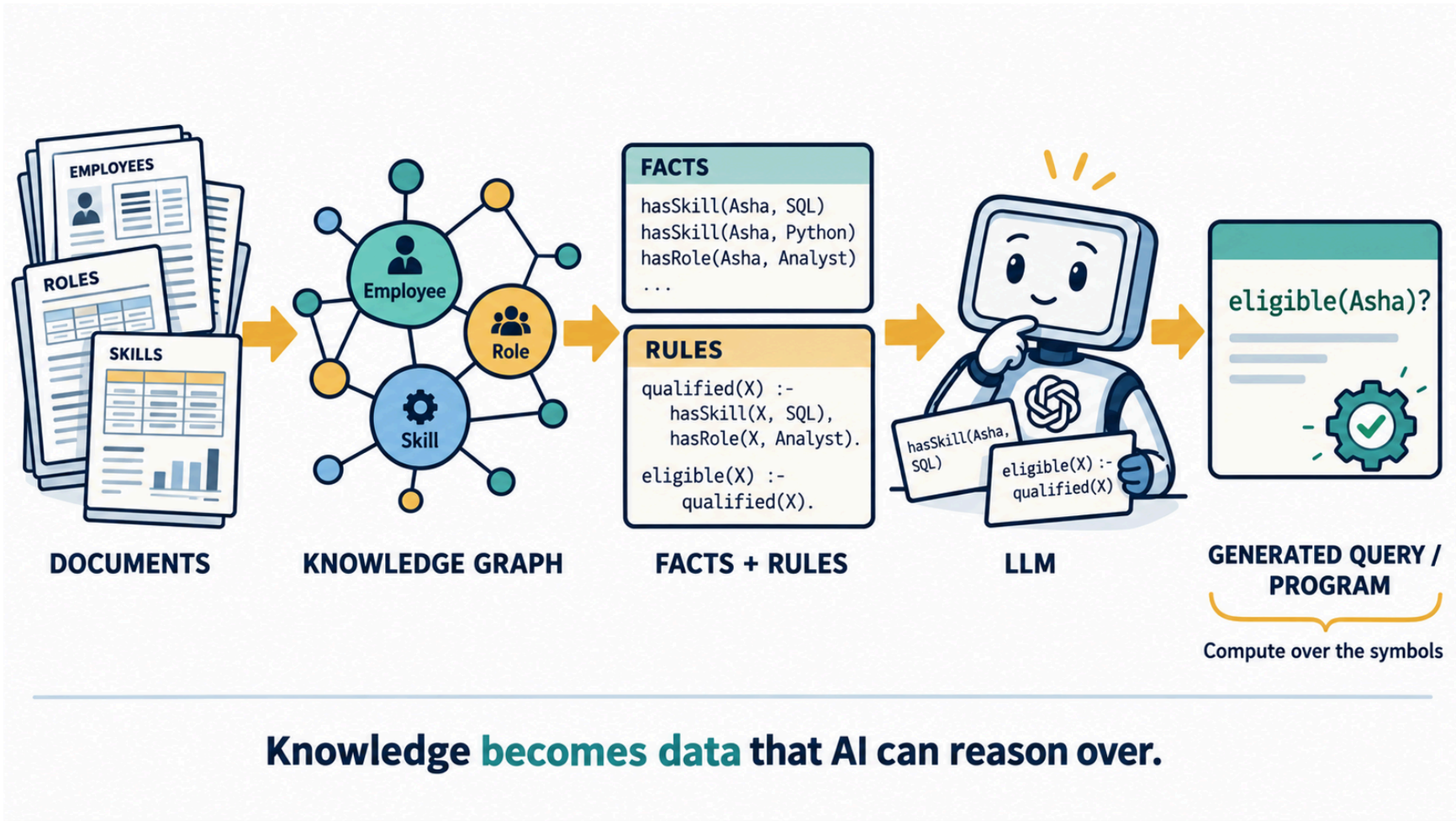
Facts, ontologies, and rules are also data.

Knowledge engineering makes their meaning explicit. The model can produce a query or program that reasons over these symbols; a rule engine computes the result.

The aim: move from generating an answer to computing it from stated facts and rules.

© 2026 Soma S Dhavala

Turn knowledge into symbols that AI can compute over



Models can generate queries and programs that reason over our knowledge.

© 2026 Soma S Dhavala

Valuation

To value data, ask how it helps someone achieve a result.

© 2026 Soma S Dhavala

What data costs to collect is different from what it is worth

COST	Producer: expenditure incurred
VALUE	Buyer: expected or realized benefit
PRICE	Transaction: what the buyer pays

Hypothetical corpus: 1,000 hours of Kashmiri speech

Coyle and Manley, What is the value of data? A review of empirical methods, 2024

Cost belongs to production; value depends on the buyer's use.

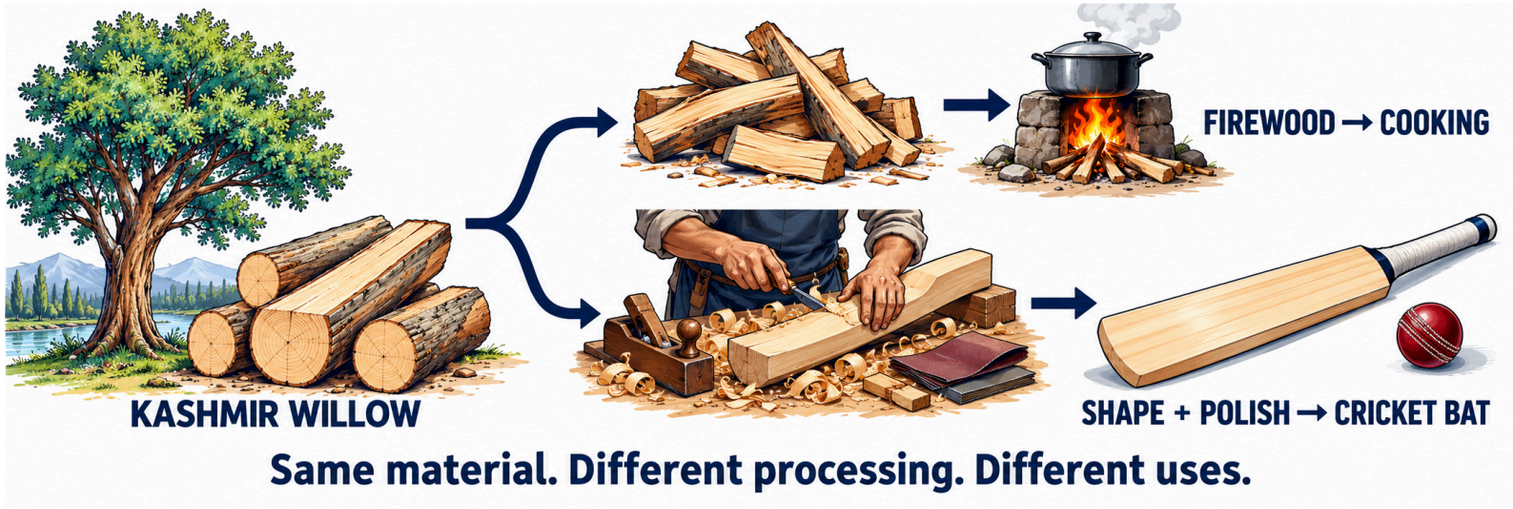
Cost: what the producer spends.

Price: what the buyer pays in a transaction.

Value: the benefit the buyer expects or realizes from using it.

© 2026 Soma S Dhavala

The same raw material can create different value in different uses



Value depends on the use, the user, and the processing.

Firewood provides heat. A crafted bat provides playing performance.

Processing costs differ. A batter may pay a premium for suitable wood and craftsmanship.

Data also matters through what it adds to a particular task.

DATA FOLLOWS THE SAME LOGIC

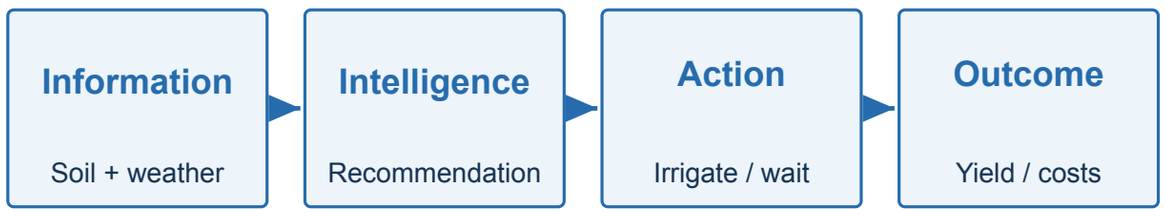


Government of India — willow cricket bats · ABC/AP — willow's firewood and bat uses

© 2026 Soma S Dhavala

Data creates value through an ecosystem that enables action





Coyle and Manley, 2024

Value depends on whether the farmer can act on the advice.

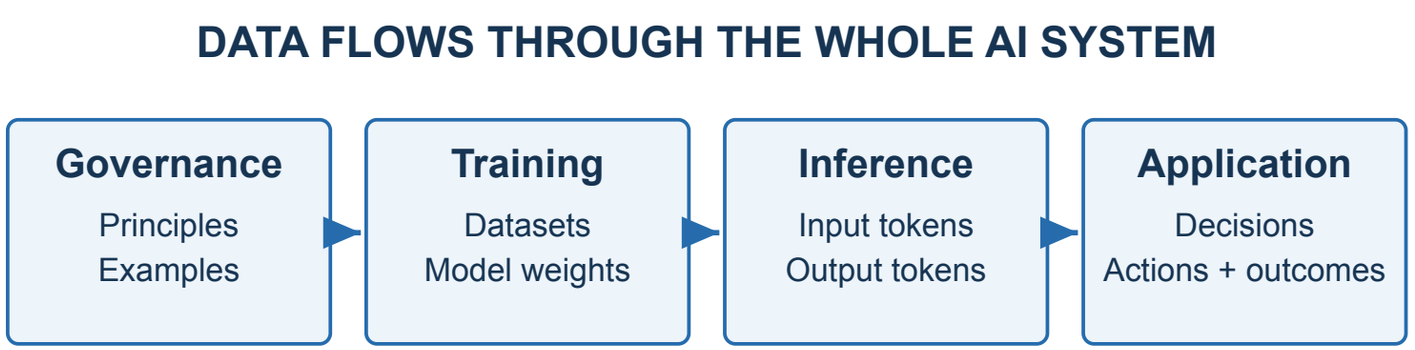
Soil and weather data guide irrigation. Water, pumps, power, labor, and timely services make action possible.

Value appears in outcomes: saved water, lower costs, or a better harvest.

If it is not actionable, it has zero value.

© 2026 Soma S Dhavala

Cost, price, and value follow data through the AI system



Coyle and Manley, 2024 · Amodei–Patel interview, 2026

Training data and generated outputs both have cost, price, and value.

Tokens meter input and output consumption at inference. The same token volume can support very different benefits.

Follow the data into the application to see what it enables.

© 2026 Soma S Dhavala

Treat reusable training data as an investment in AI infrastructure

CAPEX LENS

Build a reusable data asset

Acquire rights + curate

Document + validate

Train and improve models

Reuse across projects

Benefits across future periods

OPEX LENS

Consume an ongoing service

Live data feeds + access

Inference input/output tokens

Operate applications

Pay for continuing usage

Benefits from current use

Fund durable datasets as part of foundation-model infrastructure.

Data can build lasting capability or support current operations.

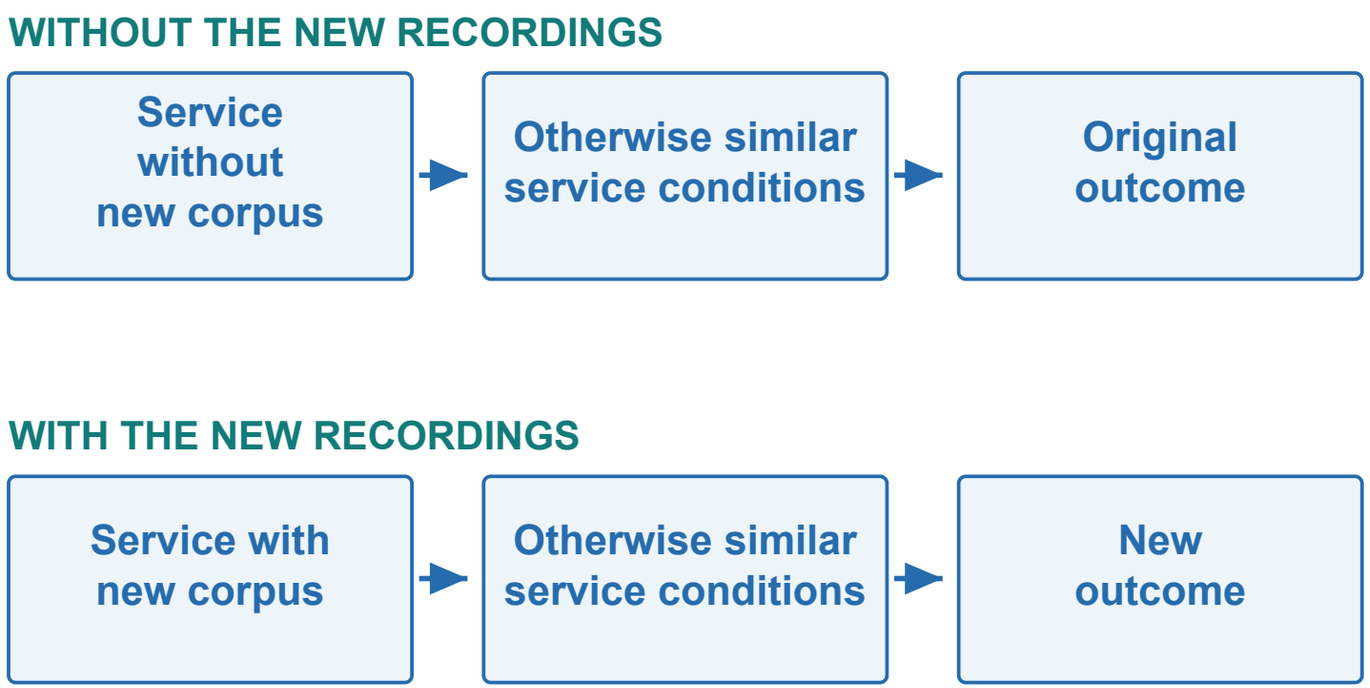
A capital-investment mindset supports multi-year funding for datasets, rights, quality, and stewardship.

This can help make foundation-model development a sustained programme.

Economic investment lens; accounting capitalization depends on applicable recognition rules. OECD — data as a produced asset · IAS 38 — intangible assets

© 2026 Soma S Dhavala

Compare results with and without the new data



Ask what improved because this dataset was added.

Compare the speech service under similar conditions, with and without the recordings. Separate the data's contribution from changes to the model or service.

Ghorbani and Zou, Data Shapley, ICML 2019

© 2026 Soma S Dhavala

Choose a valuation method that answers your question

Budget production	Cost / replacement approach
Estimate use value	Expected net benefit approach
Negotiate license	Market comparables + rights
Support public access	Social and non-monetary benefit

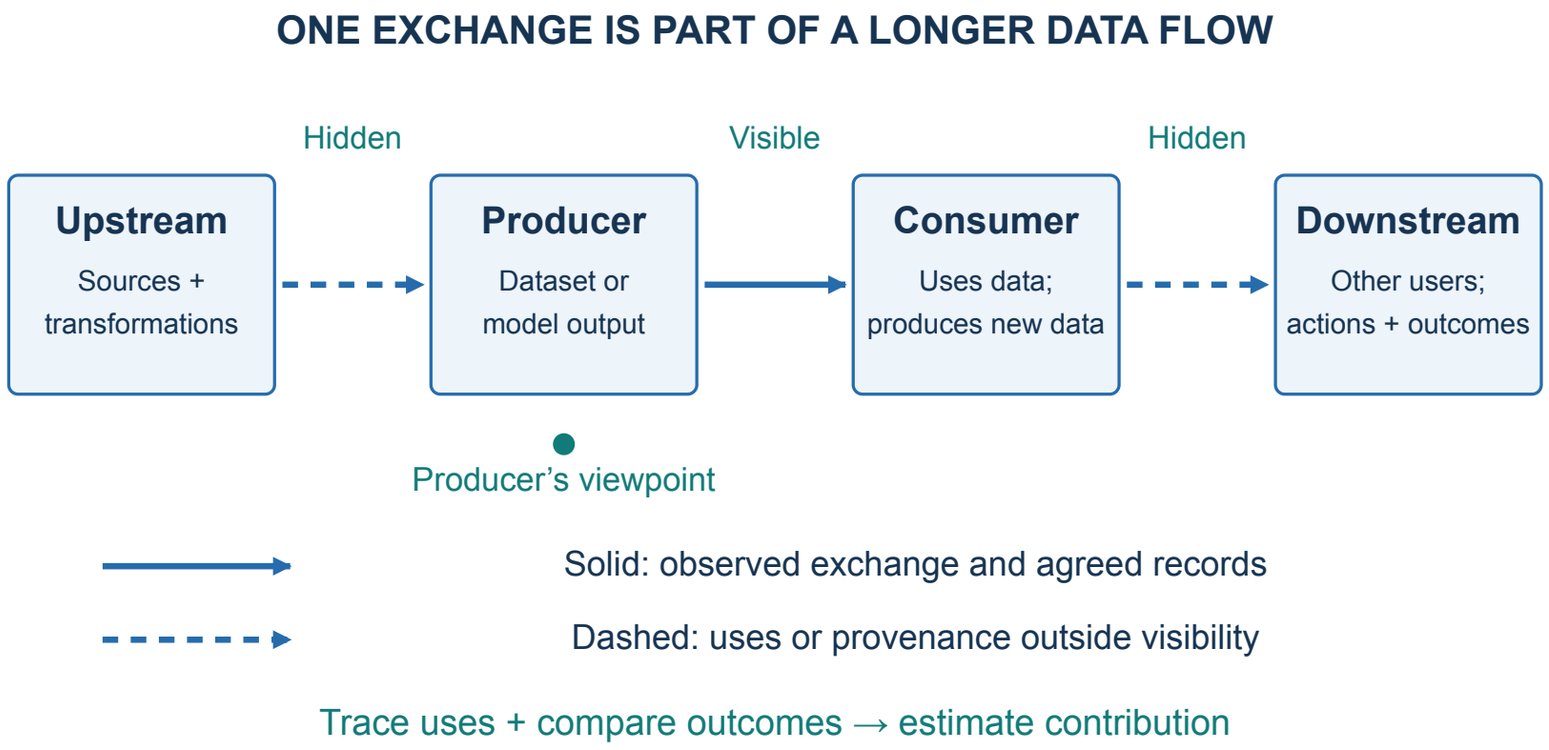
Decide what you want to estimate first.

Collection costs help plan a budget. Comparable deals help negotiate a licence. Expected improvements help estimate benefits to a service or community.

Coyle and Manley, What is the value of data? A review of empirical methods, 2024 · Fleckenstein et al., A Review of Data Valuation Approaches, 2023

© 2026 Soma S Dhavala

A producer sees one exchange, while value can arise further downstream



A consumer can become the next producer: a dataset supports a model, whose outputs support another service.

Coyle and Manley, What is the value of data? A review of empirical methods, 2024

Attribution needs visibility beyond the immediate buyer.

A producer sees its costs and the sale. Later transformations and benefits may be hidden.

Track relevant uses and compare outcomes. Where evidence is missing, make the assumptions and uncertainty explicit.

© 2026 Soma S Dhavala

Explain how you arrived at the data's value

Scope	Service • user • permissions • time period
Comparison	Results with and without the new data
Net benefit	Benefits • costs • risks
Uncertainty	Range • assumptions • what is unknown

Record production cost and negotiated price separately

Coyle & Manley, 2024 — valuation methods · Fleckenstein et al., 2023 — valuation framework

Show the calculation and the assumptions behind it.

For the speech recordings, name the service, expected improvement, costs, and time period. Show how the estimate changes if the assumptions change.

© 2026 Soma S Dhavala

Data is hard to value because its benefits appear later

COST	Money spent collecting and preparing data
PRICE	What a buyer pays to use the data
VALUE	How much the data helps a particular use
MISSING EVIDENCE	Use → action → outcome → comparison

The same data can have different value in different uses

Cab surge pricing: the same route can have a higher fare when demand exceeds available drivers. Price responds to scarcity; value also depends on the passenger's purpose and urgency.

Uber — how surge pricing responds to supply and demand

The difficult part is finding how much the data helped.

Collection costs and sale prices can be recorded. Estimating value also needs evidence of results and a comparison with what would have happened without the data.

© 2026 Soma S Dhavala

Data is a hard problem at every stage of AI

Stage of the AI system	What the data must support
Governing principles	Record whose views shaped the rules, what was agreed, and how those rules can change.
Collection and training	Provide examples that teach the model how to apply the rules, including exceptions and corrected mistakes.
Use and outcomes	Record how the data was used and what changed, so we can estimate its contribution to a result.
Reasoning and revision	Represent knowledge as facts and rules with sources. Keep the evidence needed to check an answer or correct a decision.

> **Data is not a one-time input to training.**
>
> Every stage creates new data requirements—and needs methods that can use that data.

© 2026 Soma S Dhavala

Better AI requires us to rethink the data

What each part tells us	What we need to build
Values: start with the governing principles. Alignment begins before training and continues through appeals and revision.	Data that records whose values were adopted, teaches those principles through examples, and supports correction when decisions go wrong.
Verification: make knowledge computable. Models can generate queries and programs that reason over explicit symbols.	Facts, relationships, ontologies, and rules—linked to their sources and usable by software that computes and checks results.
Valuation: assess data throughout the system. Input data, training examples, intermediate representations, and generated outputs can each contribute value.	For each use, identify who benefits and compare outcomes with an alternative. Track costs and prices separately from the value contributed.

> **Data is a hard problem across the whole system.**
>
> We need to design the data and the methods that use it together—from governing principles to real-world outcomes.

© 2026 Soma S Dhavala