Epistemic Due Diligence (EDD)

Mapping Invisible Civilisational Risk and Epistemic Narrowing in Foundation Models

R.T. van Vroonhoven

Author

Renske Therese van Vroonhoven

This whitepaper was developed as part of ongoing research with the University of Cambridge, and during the Thesis Accelerator Fellowship with Effective Thesis.

My sincere gratitude goes out to all advisors and collaborators involved in this project.

Index

Executive Summary

Modern institutions depend on legibility: simplifying complex realities into standardised categories, metrics, and records in order to coordinate action at scale. These processes create ***epistemic filters*** that shape what can count as evidence. Knowledge that is quantifiable, codifiable, and institutionally recognisable passes easily, while tacit, embodied, local, and relational forms of understanding often remain invisible or heavily burdened by translation requirements.

Foundation models intensify these dynamics. Trained on vast but historically filtered digital corpora, they inherit the biases of what has been written, digitised, preserved, and made machine-readable. As these systems increasingly mediate search, classification, summarisation, and decision-making across domains, they risk consolidating existing epistemic filters into shared infrastructure. ***The result is epistemic narrowing:*** institutional knowledge systems becoming progressively more dependent on patterns that are legible to data and models, while forms of knowledge that resist codification remain systematically excluded.

This paper hypothesises Epistemic Due Diligence (EDD) as a tentative framework for identifying and governing these risks. EDD focuses attention on the upstream design of epistemic infrastructures by asking four diagnostic questions:

1. *Whose knowledge is excluded or heavily burdened by existing systems?*
2. *What forms of knowledge remain outside the frame entirely?*
3. *Who has authority to define what counts as evidence?*
4. *Which decisions should remain revisable under conditions of epistemic uncertainty?*

Applied across data design, model development, deployment, and evaluation, ***these questions help institutions treat absences not merely as missing data but as potential signals of structural exclusion.*** By making epistemic filters visible and contestable, Epistemic Due Diligence aims to support knowledge systems that remain accountable to what they cannot easily capture.

Glossary

Burden of legibility

The unevenly distributed epistemic labour required to make oneself, one's experience, or one's knowledge recognisable within institutional systems. It includes learning institutional categories, translating experience to fit them, suppressing what does not fit, and performing credibility in the face of scepticism. The burden falls heaviest on those whose ways of knowing are furthest from institutional defaults.

Epistemic due diligence (EDD)

A systematic practice of identifying what is missing from knowledge systems, whose voices remain unheard, and which forms of understanding resist capture in existing data structures. EDD asks four diagnostic questions: Whose knowledge is excluded or burdened? What lies outside the frame? Who decides what counts? Which decisions should remain revisable?

Epistemic filter

The representational mechanisms through which institutions determine what enters as evidence, what becomes actionable, and what is registered as absence. Filters operate at multiple levels: attention, recording, categorisation, quantification, archiving. They are necessary for institutional action but never neutral.

Epistemic narrowing

The progressive contraction of what counts as knowledge within institutional systems, as epistemic filters compound across domains and harden into infrastructure. Foundation models intensify this dynamic by consolidating filters that were previously distributed across different institutions.

Legibility bias

The systematic overrepresentation, in training corpora and institutional records, of knowledge that is written, standardised, digitised, preserved, and expressed in dominant languages. Legibility bias is not random but shaped by historical patterns of power, resources, and institutional attention.

Organised epistemic irresponsibility

A governance condition in which responsibility for epistemic exclusion is distributed across multiple actors—dataset creators, model developers, deployers, users—such that no single actor can be held accountable, even as harms compound. The normal operation of layered systems produces exclusion without anyone owning it.

Structural inadmissibility

The condition of knowledge that cannot enter institutional systems without fundamental transformation or destruction. Not merely difficult to capture, but impossible to capture in institutional terms while remaining what it is. Examples include tacit skill, oral traditions that depend on oral transmission, embodied knowledge, spiritual experience, and relational goods.

Testimomial smothering

When speakers truncate their testimony because they perceive that their audience is unable or unwilling to give it proper uptake. A form of silencing that operates even when speakers have credibility and conceptual resources, because the costs of speaking are too high or the likelihood of being heard too low.

This page intentionally left blanc

Chapter 1

Introduction

The Wald Problem - Institutional Filters - Epistemic Crisis - Tacit Knowledge - Epistemic Injustice - Foundation Models - Epistemic Due Diligence Introduced

Chapter 1

Introduction

In 1943, the statistician Abraham Wald was confronted with a life-or-death problem. Allied aircrafts were being lost over Europe at an alarming rate, and military planners needed to know where to add armour for maximum protection with minimal additional weight. Analysts examined the damage on returning aircraft and observed that bullet holes were concentrated on the wings and fuselage. Their instinct was clear: reinforce the areas that appeared to be hit most often.

Wald argued the opposite.

He realised they were looking at the wrong data—the wrong planes. The fact that aircraft shot in the wings and fuselage made it back showed that those areas could withstand damage. The planes that did not return were likely hit elsewhere—in the engine or cockpit, areas that appeared pristine on the survivors. Reinforce the untouched areas, Wald insisted. He became the textbook example of survivorship bias (Wald, 1943; Mangel & Samaniego, 1984).

The lesson extends far beyond wartime aviation. It reveals a fundamental flaw in how we—and the systems we build—come to know the world: we infer from what reaches us and disregard what we do not see. What we observe is only what has survived some filter: time, selection, attention, institutional processing, linguistic expression. The problem is not simply missing data. As Wald's case demonstrates, absences are often not empty at all. They are full of information.

This is, at its heart, an epistemological problem. Epistemology is the study of how we know things. And today, that problem has become urgent in new ways. Many scholars describe a crisis in how knowledge is produced and validated, though the terms they use vary depending on what they take to be the core difficulty.

One of the most familiar terms is post-truth, declared Word of the Year by the Oxford English Dictionary in 2016 (Oxford Languages, 2016). The term describes circumstances in which emotion and personal conviction take precedence over shared knowledge, reason, and objective fact in shaping public opinion. Its rise coincided with Britain's EU referendum and the first election of Donald Trump—moments when the status of truth itself appeared newly contested.

But the crisis runs deeper. Thomas Zoglauer (2022) frames the problem as a "truth crisis" in which echo chambers, conspiracy theories, and fake news increasingly blur the boundaries between truth and lies, knowledge and opinion. Communication researchers Neuberger et al. (2022) identify an epistemic crisis in democratic debate, visible in the fragmentation of the "traditional linear knowledge order": journalism once served as the central intermediary between societal sources and audiences; now digital platforms enable direct connections that merge knowledge production, verification, and dissemination.

Robin McKenna (2023) critiques traditional epistemology for relying on idealised assumptions that obscure real-world social dynamics. Quassim Cassam (2019) develops a "vice epistemology" to analyse intellectual flaws such as closed-mindedness and arrogance, which thrive in modern conditions, while Ian James Kidd (2019) introduces "epistemic corruption" to describe how institutional incentives foster such failings.

Studies of QAnon and COVID-19 conspiracies have been interpreted through Habermas's (1975) legitimation crisis and Laudan's (1984) epistemic crisis, highlighting how declining institutional trust destabilises shared standards of justification.

These analyses converge on a common diagnosis: the problem extends beyond misinformation to the erosion of the institutional infrastructures that sustain knowledge—trusted intermediaries, reliable records, and recognised epistemic authorities.

For our purposes, we can synthesise these perspectives into a working definition:

An epistemic crisis occurs when the foundational mechanisms for distinguishing knowledge from noise—institutional filters, reliable records, audible voices—erode under suspicion, rendering the status of truth more contested than in stable conditions.

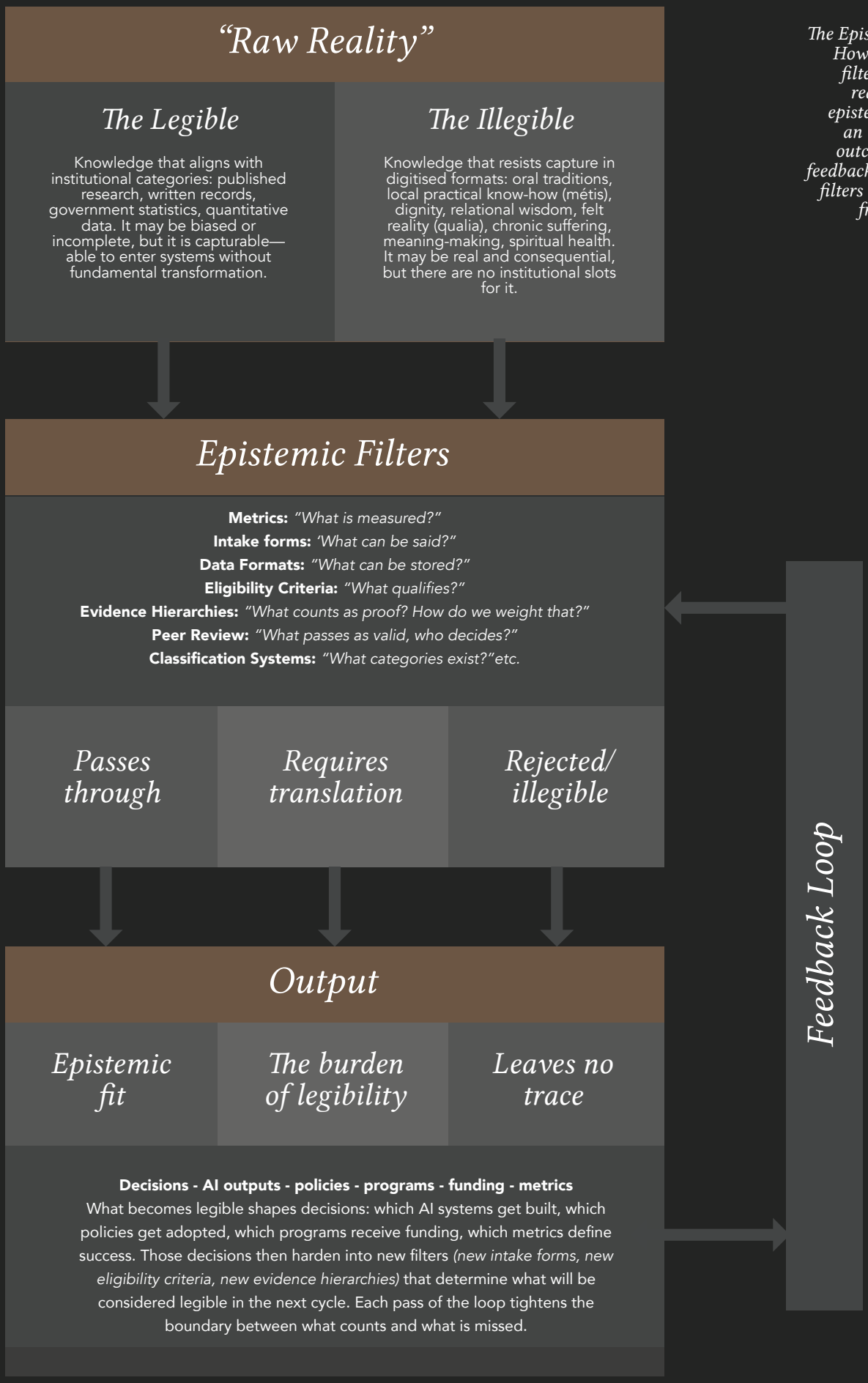


Figure 1: The Epistemic Process How knowledge is filtered from raw reality, through epistemic filters, to an outcome. This outcome creates a feedback loop, as new filters are generated from this data.

But the problem is not only about contested truth. It is also about what can never become truth-claims at all. This invites a deeper ontological debate: not everything that is true or knowable can be captured, measured, or recorded. Wittgenstein's (1922) insight that "not all that is the case can be pictured by true propositions" reveals the limits of propositional knowledge. Much of what we know remains tacit (Polanyi, 1966), embodied in skills and intuitions that resist articulation (Merleau-Ponty, 1962), irreducible to language (Levinas, 1969), or ineffable (Jonas, 1984), evading linguistic expression altogether.

These are not merely overlooked data points. They are structural blind spots in how knowledge can be gathered at all. The spaces where language falters or fails in written records often hold vital realities: unarticulated needs, unvoiced dissent, pre-linguistic perceptions (the body's sense of balance, the infant's recognition of care), qualia (like the taste of ice cream or the sensation of cold), intrinsic values (such as dignity), and relational goods (like love and belonging).

And these blind spots are not neutral. They intersect with social power in ways that philosophy has recently learned to name. Miranda Fricker (2007), in Epistemic Injustice, identifies two distinct wrongs done to people specifically in their capacity as knowers. Testimonial injustice occurs when prejudice leads us to doubt a speaker's credibility: the woman not believed about her pain, the community elder dismissed as merely anecdotal, the local practitioner whose expertise cannot register against "evidence-based" standards. Hermeneutical injustice occurs when gaps in our shared interpretive resources leave people unable to make sense of their own experiences: the sufferer of workplace harassment before the concept existed, the patient whose symptoms do not fit diagnostic categories, the community whose forms of life have no name in the language of policy.

Both forms converge on a single point: the burden of becoming legible does not fall equally. It falls heaviest on those whose ways of knowing are already devalued, whose categories are already marginal, whose voices must strain to be heard through filters not of their making. When knowledge cannot be articulated because language itself falters, the resulting silence is not neutral. It is structured by who has power to define what counts as knowledge in the first place.

Now consider the technology that increasingly mediates our institutions. Foundation models—large language models and their relatives—are trained on vast but survivorship-filtered corpora that prioritise linguistically expressible, digitally recorded content. They excel at propositional patterns yet systematically undervalue tacit, embodied, and ineffable knowledge.

Scholars have warned that such systems may reproduce or amplify existing epistemic limitations, privileging what can be recorded while marginalising what remains tacit or under-documented (Bender et al., 2021; Floridi et al., 2018). This is epistemic narrowing amplified to civilisational scale. And it creates a self-reinforcing feedback loop: what becomes legible shapes which AI systems are built, which policies adopted, which programs funded—which then harden into new filters that determine what will be considered legible in the next cycle.

The stakes can be stated simply:

If... the things that cannot be written down are systematically disregarded, we risk experiencing our lives as less meaningful.

If... institutions appear oblivious to what makes life meaningful, how can we see them as legitimate?

If... we outsource core decisions to AI systems trained solely on legible data, we amplify these blind spots, scaling survivorship bias into civilisational risk, as systems optimise for the recorded while missing the vital unrecorded.

And... because what those systems output shapes future policies, funding, and metrics—which in turn determine what data gets collected next—the cycle tightens with each pass, locking out the illegible ever more completely.

This paper proposes a framework for Epistemic Due Diligence: a systematic practice for identifying what is missing from knowledge systems, whose voices remain unheard, and which forms of understanding resist capture in existing data structures.

Drawing on Wald's insight, and engaging with the philosophical traditions outlined above, we aim to turn survivorship bias from a liability into a diagnostic tool—a way of reading absences for the information they contain. The question is not whether our knowledge is incomplete. It always is. The question is whether we have the tools to see what we are missing, and to hear those who have been silenced.

Chapter 2

Legibility, Simplification, and Epistemic Filters

Institutions and Representations - Scott: Legibility - Bowker and Star: Classificatios - Porter: Quantification - Defining Epistemic Filters

Chapter 2

Legibility, Simplification, and Epistemic Filters

Institutional action depends on representations. A state cannot manage a forest without maps that simplify infinite complexity into tractable units; a hospital cannot treat patients without medical records that translate embodied distress into standardised codes; a school cannot educate without grades that compress learning into comparable scores. As Scott (1998) demonstrated, modern governance depends on making societies legible—rendering complex, local, and practical knowledge into forms that permit centralised administration. The cadastral survey, the census, the standardised metric system: each is a technology of legibility that enables intervention at scale. To govern is to simplify.

This simplification is not merely a necessary evil but a positive condition of institutional action. States and large organisations require standardisation to compare, aggregate, and intervene across diverse contexts. Without legibility, there can be no public health policy, no economic planning, no equitable distribution of resources. Scott distinguishes between the necessity of simplification and the hubris of believing that simplified representations capture reality in full. The problem is not that states simplify—they must—but that they often forget they are simplifying. They mistake the map for the territory, the metric for the phenomenon, the category for the case.

But simplification does not merely reduce complexity; it also selects which aspects of reality matter. The map that shows roads and borders does not show soil quality or local place-names; the medical record that codes diagnoses does not capture the patient's experience of illness; the educational assessment that measures numeracy does not register creativity or practical wisdom.

The "Legible" World

Published Research - Government Statistics - Digitised Texts - Structured Data - Social Media Posts - Written Accounts

Epistemic Filtering

What does it take for knowledge to show up 'above the waterline'?

The "Illegible" World

Tacit Knowledge (Polanyi), Oral Traditions, Embodied Experience (Merleau-Ponty), Locally Entangled Knowledges, Métis (Scott), Relational Goods, Qualia (the feeling of pain), Dignity, Unarticulated Needs, the sense of something (an experience, life) "having meaning", Narrative Cohesion.

These omissions are not random but patterned by what Bowker and Star (1999) call classification systems—the often-invisible infrastructures that organise how we sort, store, and retrieve information. Classification systems are not passive descriptions of pre-existing categories; they actively shape what becomes visible, countable, and actionable. The International Classification of Diseases determines which conditions exist for purposes of healthcare funding; racial and ethnic categories determine which groups can be counted for anti-discrimination policy. Once embedded in infrastructure, these classifications become naturalised, appearing not as human decisions but as features of reality itself.

Theodore Porter (1995) adds a crucial further dimension in Trust in Numbers. Quantification, he argues, often emerges not primarily from truth-seeking but from bureaucratic and political demands for objectivity, impersonality, and legitimacy. When organisations face distrust or must coordinate across distances, they turn to numbers as a technology of trust—a way of making decisions that appear mechanical rather than discretionary, rule-bound rather than subjective. This mechanical objectivity has real virtues: it can constrain bias, enable comparison, and create accountability. But it also privileges what can be quantified over what cannot, and what can be standardised over what resists standardisation. The very features that make quantification attractive to institutions—its impersonality, its seeming freedom from local context—are also the features that make it systematically blind to forms of knowledge that depend on context, judgment, and qualitative assessment.

From these sources, we can derive a working concept: epistemic filters are the representational mechanisms through which institutions determine what enters as evidence, what becomes actionable, and what is registered as absence. Epistemic filters operate at multiple levels: at the level of what is noticed versus what passes unremarked; at the level of what is recorded versus what remains oral or ephemeral; at the level of what is categorised versus what resists classification; at the level of what is quantified versus what is described; at the level of what enters databases versus what remains in local archives or individual memory. These filters are not added after the fact of knowledge production but are constitutive of it: they shape what can be known in institutional terms at all. To speak of epistemic filters is not to complain about imperfection but to name the unavoidable condition of institutional epistemology: institutions know through simplification, and simplification always selects.

Implication	*What is Lost*	*Filter*
Normative (judgement: what matters)	What passes unremarked	Attention
Cumulative (gone forever)	Oral, ephemeral, informal	Recording
Normative (judgement: what fits?	What resists classification	Categorisation
Infrastructural (Embedded in metrics)	Qualitative, contextual	Quantification
Cumulative & infrastructural	Local memory, non-digitised	Archiving

Figure 2: The Epistemic Filter Stack
How institutional knowledge is produced through successive layers of selection, and what is lost at each stage.

This concept has three crucial implications. First, filters are not merely technical but normative: they embody judgments about what matters, what is real, what deserves attention. Second, filters are infrastructural: once embedded in forms, databases, and routines, they shape countless decisions without requiring explicit re-authorisation. Third, filters are cumulative: what is filtered out at one stage cannot be recovered later, so early decisions about what to record or how to categorise propagate through entire knowledge systems. Understanding these filters is not a prelude to eliminating them—an impossibility—but to making them visible, contestable, and accountable.

Chapter 3

The Burden of Legibility

From Administrative Burden to Epistemic Labour - Performing Credibility - Uneven Distribution - Formal Definition

Chapter 3

The Burden of Legibility

If institutions know through simplification, subjects must often do work to become legible to them. What must individuals, communities, and forms of knowledge do to become visible within institutional systems and pass through epistemic filters?

A useful precedent exists in the public policy literature on administrative burden. Moynihan, Herd, and Harvey (2015) define administrative burden as the learning, psychological, and compliance costs that individuals experience in interacting with government programs. Herd and Moynihan (2019) subsequently showed that these burdens are not merely incidental but are often designed into programs in ways that systematically exclude eligible recipients—a form of what they term "administrative exclusion." The administrative burden literature demonstrates that institutions routinely externalise labour onto citizens, requiring them to navigate complex requirements, produce documentation, and conform to bureaucratic expectations.

I use the term *"burden of legibility"* to name the unevenly distributed epistemic labour required to make persons, experiences, or knowledges admissible within institutional systems. Where administrative burden focuses on practical costs, the burden of legibility concerns the interpretive and translational work required to make one's experience recognisable in institutional terms.

It is not enough to be in need; one must render that need into categories that qualify for assistance. It is not enough to be in pain; one must describe that pain in terms that fit diagnostic criteria. It is not enough to possess local knowledge; one must translate that knowledge into indicators and logframes that funders recognise. This is labour—often invisible, often uncompensated, and disproportionately borne by those whose experiences lie furthest from institutional categories.

ITN-Analysis

Importance

Upstream intervention before lock-in multiplies effect. This makes it an important problem to solve, as it not only effects many people, but it can be solved upstream, meaning a small intervention has a high impact potential.
Once epistemic filters harden into infrastructure, switching costs make revision prohibitive. Preserving epistemic diversity prevents correlated blind spots across healthcare, education, and social services, blind spots that could cascade into systemic failure.
This is civilisational-scale risk prevention.

Tractability

Concrete interventions exist. Four diagnostic questions can be integrated into procurement standards, dataset documentation, and model development practices today. Review panels, community governance mechanisms, and revision triggers are institutionally feasible—they don't require solving AI alignment first.

Neglectedness

AI governance today fixates on alignment, safety, and demographic bias—all of which assume the relevant knowledge is already captured. Epistemic infrastructure—what counts as evidence in the first place—remains almost entirely off the radar because it's invisible by design; you don't notice a classification system until it breaks, and by then exclusion is naturalised. The problem falls between disciplinary stools: computer scientists own models, philosophers own epistemology, policymakers own regulation—no one owns the intersection.

Clear causal pathway

Visibility → Contestation → Redesign → Legitimacy.
Making filters visible enables affected communities to challenge them. Challenge creates pressure for redesign. Redesign that responds to exclusion rebuilds institutional legitimacy. The pathway is concrete and observable.

Robust to uncertainty

EDD creates adaptive capacity regardless of which specific knowledge forms turn out to matter most. You don't need to predict the future; you need institutions capable of revising themselves when blind spots become apparent. That's resilience, not prediction.

Figure 1: The Epistemic Process How knowledge is filtered from raw reality, through epistemic filters, to an outcome. This outcome creates a feedback loop, as new filters are generated from this data.

Healthcare provides a vivid illustration. Werner and Malterud (2003) studied women with chronic pain whose experiences were systematically disbelieved or minimised by medical professionals. Their title captures the phenomenon: *"It is hard work behaving as a credible patient."* These women had to perform credibility—presenting themselves in ways that would be taken seriously, narrating their symptoms in medically legible terms, and suppressing aspects of their experience that did not fit biomedical frameworks.

Malterud's (2000) work on medically unexplained symptoms shows how patients whose conditions lack clear pathological correlates must work even harder to be heard. Havi Carel's (2016) phenomenological analysis further demonstrates how illness disrupts the taken-for-granted relationship between body and world, and how medical institutions often fail to register this existential dimension. The issue is not merely whether patients are believed, but whether their experience can enter the system in a form the system is built to recognise.

The same logic operates across domains. In international development, communities must translate local realities into the categories required by donor agencies: logframes, indicators, and results matrices. Robert Chambers (1997) critiqued how development professionals' reliance on structured methods and pre-set categories systematically excluded local knowledge that did not fit.

Rosalind Eyben (2013) likewise shows how aid relationships are structured by what can be counted and reported, rendering invisible the relational work and local improvisation that often determine project outcomes. Indigenous communities seeking recognition of land rights must translate ancestral relationships into the property categories of colonial legal systems. Farmers seeking agricultural extension services must translate practical know-how into technical language. In each case, the burden falls on those seeking recognition to make themselves legible in terms they did not design.

Crucially, this burden is not equally distributed. Dominant groups often appear naturally legible to institutions because institutional categories were historically designed with their experiences in mind. The white, male, middle-class patient is more likely to find his symptoms believed; the native English speaker navigates bureaucratic forms more easily; the formally educated applicant recognises categories and completes them without strain.

For marginalised groups, legibility requires translation, and translation costs are higher. Fricker's (2007) analysis of testimonial injustice identifies one mechanism: when speakers face credibility deficits due to identity prejudice, they must work harder to be believed, and may never succeed regardless of effort. The burden of legibility is therefore not merely a practical inconvenience but a site of potential injustice—a systematic unevenness in what it costs to be heard.

The burden of legibility can thus be understood as unequally distributed epistemic labour: the work of learning institutional categories, translating experience into them, suppressing what does not fit, performing credibility in the face of scepticism, and persisting through repeated failure. Recognising this burden does not imply that institutions should eliminate all translation requirements—some standardisation is necessary for coordinated action—but it does require asking who bears the costs of legibility and whether those costs are justified.

Chapter 4

Epistemic Injustice and Structural Inadmissibility

Testimonial and Hermeneutical Injustice - Three Categories of Knowledge - Structural Inadmissibility - Institutional Absence

Chapter 4

Epistemic Injustice and Structural Inadmissibility

When the burden of legibility becomes too heavy, or when translation fails entirely, the result is not merely inconvenience but a wrong done to people in their capacity as knowers. Fricker's (2007) analysis provides the foundational vocabulary.

Testimonial injustice occurs when prejudice causes a hearer to assign deflated credibility to a speaker's word. The woman whose pain is dismissed as emotional rather than physical; the Black witness whose testimony is discounted by a jury; the community elder whose oral history is treated as mere anecdote—each suffers a distinctively epistemic wrong. They are wronged in their capacity as givers of knowledge, degraded as knowers.

Hermeneutical injustice occurs when a gap in collective interpretive resources leaves someone unable to make sense of an experience that is significantly harmful to them. The woman sexually harassed before the term existed, who lacked conceptual resources to understand what was happening; the sufferer of chronic fatigue syndrome whose symptoms do not fit available diagnostic categories; the indigenous community whose forms of spiritual relationship have no name in the language of property law—each suffers a structural disadvantage in making their experience intelligible, even to themselves. Hermeneutical injustice points to gaps in shared interpretive resources that condition what can be thought, said, and understood.

Fricker's analysis can be extended. Kristie Dotson (2012, 2014) argues that epistemic oppression includes practices of silencing that operate even when speakers have both credibility and conceptual resources. Testimonial smothering occurs when speakers truncate their testimony because they perceive that their audience is unable or unwilling to give it proper uptake.

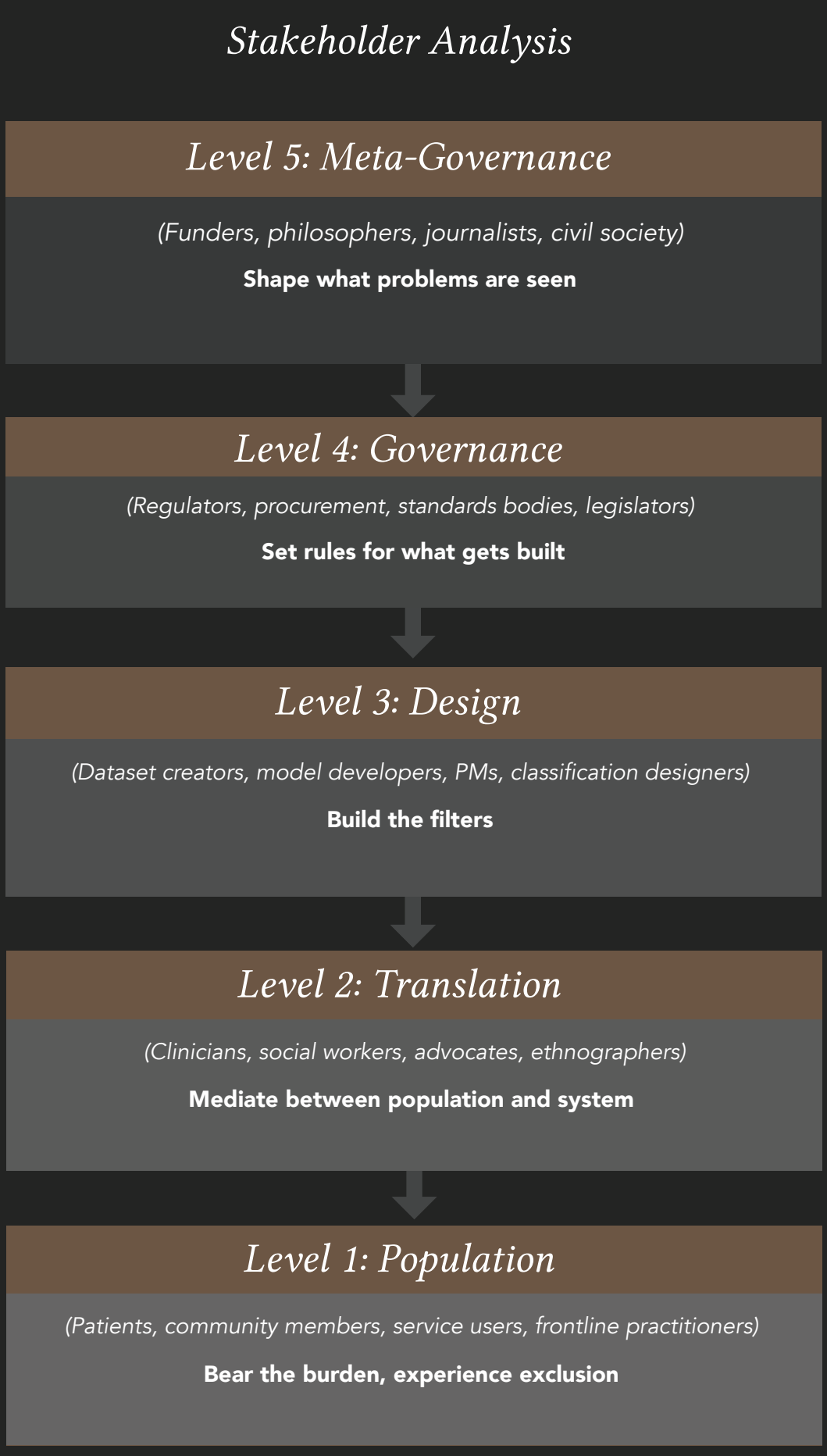


Figure 1: The Epistemic Process How knowledge is filtered from raw reality, through epistemic filters, to an outcome. This outcome creates a feedback loop, as new filters are generated from this data.

Gaile Pohlhaus (2012) develops the concept of willful hermeneutical ignorance, where dominant groups actively resist interpretive resources developed by marginalised communities, maintaining ignorance through motivated refusal. José Medina (2013) adds the importance of resistant counter-knowledges—epistemic resources that marginalised communities develop precisely because dominant institutions exclude them. Taken together, these accounts show that exclusion does not arise only from missing concepts. It also arises from failed uptake, strategic silence, and active resistance to interpretive resources developed by marginalised communities.

Building on these sources, we can distinguish three categories of knowledge in relation to institutional epistemic filters.

First, knowledge that passes easily. This is knowledge that aligns with institutional categories, fits existing forms and metrics, comes from credible speakers speaking in expected ways. It requires little translation; its bearers face low burdens of legibility. This knowledge is typically produced by dominant groups, in dominant languages, through dominant institutions, about phenomena already recognised as real and important. Its ease of passage reinforces its authority, creating a self-perpetuating cycle: what is easily legible becomes more visible, more documented, more institutionalised, and thus even more easily legible in future.

Second, knowledge that enters through costly translation. This is knowledge that requires work to become institutionally admissible—the patient who must perform credibility, the community that must produce indicators, the indigenous group that must frame land claims in property terms. This knowledge can enter, but at a cost borne by those who produce it. The cost may be practical (time, money, effort), interpretive (suppressing aspects that do not fit), or psychological (stress, stigma, alienation). This is conditional inclusion: inclusion on terms set by the institution, requiring those who seek recognition to reshape themselves to fit available categories.

Third, knowledge that is structurally inadmissible. This is knowledge that cannot enter institutional systems at all, not merely because it is difficult to translate but because there are no institutional slots for it. There are at least three reasons knowledge may be structurally inadmissible: it may resist codification without fundamental distortion; it may depend on local context that standardisation strips away; or it may lack recognised institutional categories altogether. Tacit skill—the craftworker's feel for materials, the clinician's intuitive judgment—resists codification by its nature.

Oral traditions that have never been written down, and whose authority depends on oral transmission, cannot survive translation to text. Embodied knowledge—the body's sense of balance, proprioceptive awareness—cannot be captured in propositional form. Highly local knowledge—knowledge of this place, these relationships, this community—resists the standardisation institutions require. Spiritual knowledge, relational goods, experiences of dignity and belonging may be experientially real to those who possess them yet invisible to institutional systems built on different foundations.

Structural inadmissibility is not merely a gap in data coverage but a systematic feature of institutional epistemology. Some forms of knowledge are not merely difficult to record but impossible to record without transformation that destroys what makes them knowledge. This is not a failure better data collection could remedy; it is a limit built into institutional knowledge. Recognising this limit does not mean abandoning institutions—some forms of knowledge must be sacrificed for the coordination and scale institutions enable—but it does require humility about what institutions can know and what they necessarily miss.

When institutions encounter structural inadmissibility, they do not typically register it as conflict or limitation. They register it as absence: "no data," "non-compliance," "lost to follow-up," "anecdotal," "culturally specific," "not evidence-based." These terms transform structural exclusion into administrative categories, obscuring the fact that something real was present but could not be registered. Werner and Malterud's (2003) patients were not merely difficult to diagnose; their experiences were systematically excluded by biomedical frameworks that could not accommodate them. The administrative burden literature shows how those who cannot meet compliance costs become "non-compliant" rather than "excluded by design." The epistemic injustice literature shows how those who cannot make themselves heard become "non-credible" rather than "systematically silenced." The language of institutional absence naturalises what is in fact a product of epistemic filters and their distributional consequences.

Chapter 5

Foundation Models as Epistemic Infrastructures

Learning from Traces - Legibility Bias in Training Data - General Epistemic Interfaces - Consolidating Filters

Chapter 5

Foundational Models as Epistemic Infrastructure

Foundation models learn from traces, not from the world directly. As Bender and colleagues (2021) emphasise, large language models are trained on text corpora, learning statistical regularities in language use rather than engaging the world through embodied, situated, and socially accountable forms of inquiry. They do not perceive the world directly, experience embodiment, or participate in social relationships; they model patterns in linguistic data. Everything a foundation model "knows" is mediated through texts that survived to enter its training corpus. The model's epistemology is structurally derivative of the archive from which it learns.

But training corpora are historically produced artefacts, shaped by centuries of filtering. What gets written down versus what remains oral; what gets published versus what stays in private circulation; what gets digitised versus what languishes in physical archives; what gets preserved versus what decays; what gets written in dominant languages versus what is expressed in minor or endangered languages; what platforms make visible versus what they bury; what copyright regimes permit versus what they block—each is a filter determining which texts become available for model training.

Gebru and colleagues (2021) argue for rigorous dataset documentation precisely because datasets are not neutral representations but constructed artefacts embedding the values, priorities, and blind spots of their creators. Birhane (2021) shows how these filters operate at global scale, with training data overwhelmingly dominated by English-language, Western, digitally native content, rendering invisible knowledge produced in other languages, other formats, other epistemic traditions. Noble (2018) demonstrates how search and ranking algorithms reproduce and amplify social hierarchies, privileging dominant perspectives while marginalising others.

This creates legibility bias in training corpora. Text corpora systematically privilege what is already written, standardised, platformed, and preserved. Oral traditions, embodied practices, informal knowledge, intimate communication, highly local expertise, spiritual experience—all are structurally underrepresented because they never became text, or never survived to enter the digital archive, or exist in languages and formats undervalued in dataset construction. The model learns from what is legible in textual form, and what is legible in textual form is not a random sample of human knowledge but a heavily filtered subset shaped by power, resources, and historical accident. This is epistemic narrowing: the systematic contraction of what counts as knowledge through the compounding effects of multiple filters.

As models are integrated into search and summarisation systems, they can shape what information becomes more findable, salient, and reusable. When they power summarisation, they influence what counts as salient. When they enable classification, they set categories through which phenomena are sorted. When they assist drafting, they shape what can be said and how. Foundation models thus become general epistemic interfaces—lenses through which increasing amounts of institutional and individual knowledge are filtered.

Models and closely related AI systems are increasingly used to rank candidates, process clinical or administrative text, summarise legal and educational materials, and shape access to information in platform environments. The extent of this infrastructural role varies by domain, deployment context, and governance regime, but the broader trend is toward deeper mediation of access, classification, and synthesis.

The crucial point is that foundation models do not create epistemic filters from scratch. They inherit filters that have operated for centuries: what was written down, what was preserved, what was translated into dominant languages, what was made digitally available. But they then consolidate, generalise, and routinise these filters across domains.

Where previously different institutions had different filters—hospitals used medical categories, courts used legal categories, schools used educational categories—now a single model might be used across multiple domains, embedding similar biases in healthcare, education, employment, and governance. Where previously filters were applied by humans who could exercise judgment and adjust to context, now they are applied algorithmically at scale, with the appearance of mechanical objectivity masking contingent choices embedded in training data.

A Tentative EDD Diagnostic Tool: An Exploratory Matrix for Practitioners

EDD QUESTION	*DATA CURATION*	*MODEL DEVELOPMENT*	*DEPLOYMENT PLANNING*	*MONITORING & EVALUATION*
Whose knowledge is excluded or burdened?	Which languages, dialects, or communities are under-represented in the dataset?	Does the model perform differently for groups whose knowledge required "costly translation" to be	Which user groups will face the highest "burden of legibility" to get their needs met by this system?	Are we tracking who is being served vs. excluded? Are complaints coming from certain groups?
What lies outside the frame (structural inadmissibility)?	What tacit, embodied, or relational aspects of the domain cannot be captured by this data format?	What kinds of human judgment or contextual understanding is the model incapable of learning?	What decisions are being fully automated where human contextual judgment is essential?	Are we capturing "near-misses" or cases where the system's output was technically correct but contextually inappropriate?
What decisions should remain revisable?	Are our data categories fixed, or can they evolve as we learn?	Is the model's output treated as final or as a draft/ suggestion?	Are there "human-in-the-loop" processes for high-stakes or ambiguous cases?	What is our process for updating the model or its use when a blind spot is identified?
Who defines what counts as evidence?	Were affected communities involved in defining the data schema?	Were frontline practitioners involved in defining success metrics for the model?	Do impacted communities have a formal role in oversight or governance?	Who has the authority to challenge the system's outputs and mandate changes?

And where previously the limitations of filters were at least potentially visible to those who applied them, now they are buried in billions of parameters, inaccessible to scrutiny.

This raises the possibility of epistemic narrowing at unprecedented scale. The problem is not that models are biased in the narrow sense of producing systematically unfair outcomes for particular groups—though they are, and this matters. The deeper problem is that they institutionalise a particular way of knowing that systematically privileges what can be written, standardised, and archived, while rendering tacit, embodied, oral, and local knowledge structurally invisible. Because models increasingly mediate institutional knowledge across domains, this invisibility becomes locked in, self-reinforcing, and difficult to contest.

Chapter 6

Governance Crisis: Lock-In, Cascades, and Legitimacy

Lock-In - Correlated Blind Spots - Democratic Legitimacy - Organised Epistemic Irresponsibility

Chapter 6

Governance Crisis: Lock-in, Cascades, and Legitimacy

Once epistemic filters are embedded in infrastructure, they become difficult to revise. Paul Pierson (2000) analyses path dependence in politics and institutions, showing how early decisions create increasing returns that make reversal costly. Paul David (1985) famously demonstrated this with the QWERTY keyboard, which persisted despite superior alternatives because of coordination effects and sunk investments.

By analogy with path-dependent infrastructures more generally, epistemic infrastructures may also exhibit lock-in once classifications, datasets, and model-dependent routines become embedded across organisations. Changing them requires not just recognising limitations but coordinating transitions across multiple actors, retraining personnel, revising documentation, and managing interoperability with systems that have not changed. This lock-in means epistemic filters, once established, tend to persist even when their limitations become apparent.

More concerning than isolated lock-in is the possibility of correlated blind spots. If many organisations use the same foundation models, the same benchmarks, the same training data, or the same classification systems, they can inherit the same omissions. Where previous epistemic diversity meant different institutions missed different things, epistemic consolidation means they miss the same things in the same ways.

A healthcare system, an education system, a hiring system, and a welfare system relying on the same underlying model will all be blind to the same forms of knowledge, the same experiences, the same communities. This correlation matters because these systems interact: problems invisible to healthcare may become invisible to welfare, invisible to education, invisible to employment. The person whose experience is structurally inadmissible in one domain may find themselves invisible across all domains, creating compounding exclusion.

In tightly coupled systems, one epistemic blind spot can propagate into cascading failure. Charles Perrow's (1984) analysis of normal accidents shows how complex, tightly coupled systems can fail in ways designers did not anticipate because interactions across components create novel error pathways. Pescaroli and Alexander (2015) develop this into cascading disasters, where an initial failure triggers subsequent failures through interconnected vulnerabilities.

The aim here is not to claim empirical identity between these systems, but to use established concepts of lock-in and cascade to illuminate analogous governance risks in epistemic infrastructures. A blind spot in one domain—say, a model's inability to register certain forms of local knowledge—could propagate through interconnected systems: disaster response that misses local coping strategies, public health that misses community practices, economic recovery that misses informal economies. The failure is not merely that each system operates with incomplete information, but that their interactions compound the effects of incompleteness.

These dynamics have implications for democratic legitimacy. Democratic governance depends, in part, on institutions that can register the experiences and interests of those they govern. When institutions systematically fail to see certain groups or certain forms of experience, they lose not just accuracy but legitimacy. Citizens whose knowledge is consistently excluded, whose testimony is consistently discounted, whose forms of life are consistently invisible, have little reason to regard institutions as responsive to their concerns. Epistemic exclusion thus feeds political alienation.

As theorists of epistemic democracy have argued (Warren, 2017), democratic legitimacy requires not just formal inclusion but substantive epistemic inclusion—the capacity of institutions to learn from and be responsive to the full range of affected knowledges. When epistemic filters systematically exclude marginalised perspectives, they undermine this condition.

We can name this problem organised epistemic irresponsibility. It is organised because it is produced by systems, not by individual actors acting alone. Exclusions are dispersed across forms, datasets, standards, and models, so that no single actor owns the harm. The hospital that uses a biased model did not create the bias; the tech company that produced the model did not create the training data; the archivists who curated the data did not create the historical exclusions it reflects.

Responsibility is distributed across a chain of actors, each of whom can plausibly claim they were only doing their part. And it is irresponsibility because this distribution makes accountability elusive: when everyone is partly responsible, full responsibility becomes difficult to locate and enforce. Organised epistemic irresponsibility names the governance challenge posed by layered, distributed epistemic infrastructure: the difficulty of assigning accountability when filters operate across multiple sites, when effects are indirect and delayed, and when harms are produced by normal operation rather than malfunction.

The stakes can now be restated with greater precision. Epistemic filters are necessary for institutional action, but they are not neutral. They impose uneven burdens of legibility, privileging some forms of knowledge while rendering others structurally inadmissible. Foundation models inherit, consolidate, and generalise these filters across domains, creating correlated blind spots and cascading vulnerabilities. The result is organised epistemic irresponsibility: systematic exclusion without anyone being accountable.

This is not merely a technical problem of bias in AI but a governance problem of how institutions know, whom they see, and what they miss. The question is whether we can develop practices that make these filters visible and contestable before they harden into infrastructure.

Chapter 7

Epistemic Due Diligence (EDD)

Upstream Intervention - Four Diagnostic Questions - Institutional Forms - Accountable Ignorance

Chapter 7

Epistemic Due Diligence (EDD)

If the problem is epistemic filters embedded in infrastructure, the response must operate upstream, before filters harden into unchangeable systems. Downstream fairness checks, bias audits, and algorithmic impact assessments are valuable, but they often inherit already-fixed admissibility criteria. They can tell us that a model performs differently across groups, but not whether the categories those groups are sorted into are the right categories. They can measure disparities in outcomes, but not whether the outcomes being measured capture what matters. Epistemic Due Diligence is an upstream practice for identifying exclusions, burdens, and admissibility conditions before they become embedded in infrastructure.

Epistemic Due Diligence is not a magic checklist or a one-time certification. It is a systematic practice of asking what is missing, whose voices are unheard, and which forms of understanding resist capture in existing data structures. It is due diligence because it parallels other forms of investigative care—financial due diligence, environmental due diligence, human rights due diligence—that seek to identify risks and harms before they materialise. It is epistemic because the risks and harms it addresses concern how institutions know and what they systematically fail to know. The goal is not to eliminate filtering—that is impossible—but to make filters more visible, more contestable, and more accountable.

A useful normative grounding can be found in Helen Longino's (1990) Science as Social Knowledge. Longino argues that scientific objectivity is not achieved by individual scientists purging themselves of bias but by the social character of inquiry: when multiple perspectives interact under conditions of critical discourse, they can identify assumptions and blind spots that no single perspective could see alone. Objectivity is secured not by neutrality but by structured critical interaction. Applied to epistemic infrastructure, this suggests better knowledge systems are not those claiming to be bias-free but those building in mechanisms for surfacing and contesting their own filters. Epistemic Due Diligence is one such mechanism.

Theory of Change (ToC)

Upstream Intervention

Asking four diagnostic questions during dataset design, model development, and procurement shifts attention from bias in outputs to admissibility conditions.

STS literature on path dependence (David 1985; Pierson 2000) shows early decisions create lock-in; Gebru et al. (2021) on datasheets demonstrates feasibility of structured upstream reflection

Making Filters Visible

Structured documentation of exclusions, burdens, and structural inadmissibility transforms tacit institutional assumptions into contestable objects

Bowker & Star (2000) on classification systems becoming visible only at breakdown; Longino (1990) on objectivity requiring critical interaction, not neutrality

Accountablity Mechanism

Review panels, participatory design, and revision triggers distribute responsibility and prevent organised epistemic irresponsibility

Dotson (2012, 2014) on epistemic oppression requiring structural, not merely interpersonal, remedy; Medina (2013) on epistemic friction as necessary for institutional learning

Epistemic Diversity

Institutions maintain capacity to register knowledge that resists codification, rather than consolidating around whatever is most legible

Scott (1998) on mētis and the failure of thin simplifications; Polanyi (1966) on tacit knowledge resisting articulation without loss

Fosters Legitimacy

Citizens see institutions as responsive to their forms of knowing, preventing the political alienation that follows epistemic exclusion.

Warren (2017) on epistemic dimensions of democratic legitimacy; Fricker (2007) on epistemic injustice undermining trust in knowledge systems

Figure 1: The Epistemic Process How knowledge is filtered from raw reality, through epistemic filters, to an outcome. This outcome creates a feedback loop, as new filters are generated from this data.

Key Causal Assumptions

Intervenability

Current epistemic filters are not yet fully locked in; there remains a narrow window to shape foundation model integration before path dependence forecloses alternatives.

Visibility Leads to Contestation

Making filters visible is not sufficient, but it is necessary—filters that remain invisible cannot be contested.

Diversity Prevents Harm

When filters are visible and there exist mechanisms for affected communities to challenge them, institutions can redesign systems to be more epistemically inclusive (or at least more honest about their limits).

Legitimacy Tracks

Democratic legitimacy depends not on perfect representation but on demonstrable responsiveness to what is initially invisible or excluded.

The Feedback Loop EDD aims to interrupt

CURRENT

Raw reality passes through epistemic filters, producing legible knowledge that shapes decisions
Those decisions harden into new filters—tighter categories, narrower metrics, more rigid forms
Each cycle tightens the boundary between what counts and what is missed
The system self-reinforces: what becomes legible shapes what data gets collected next, which shapes what will be legible in future

EDD INTERVENTION

Upstream questions interrupt filter design before hardening: whose knowledge excluded, what's outside frame, who decides
Ongoing monitoring tracks absences, near-misses, and ethnographic signals rather than only performance metrics
Governance mechanisms create revision triggers, community review, and participatory redesign when blind spots appear
The cycle becomes accountable rather than self-reinforcing

Four questions should guide epistemic due diligence at key decision points in the design and deployment of epistemic infrastructure.

First, whose knowledge is excluded or heavily burdened? This question directs attention to distributional effects. It asks not only whether data is representative in some statistical sense but whether the forms of knowledge required to navigate the system align with the epistemic resources of some groups more than others. It asks who must translate, who must perform credibility, who must suppress aspects of their experience to become legible.

Second, what forms of knowledge likely remain outside the current frame? This question directs attention to structural inadmissibility. It asks what forms of knowledge—tacit, embodied, oral, local, spiritual, relational—may be systematically invisible to the proposed system. It does not assume all such knowledge can or should be captured, but insists on acknowledging what is being missed.

Third, which decisions should remain revisable under deep uncertainty? This question directs attention to the temporal dimension of epistemic risk. When we cannot know what we are missing, some decisions should be made provisionally, with mechanisms for revision as blind spots become apparent. It asks what would need to be true for us to be confident in our knowledge, and what we should do if those conditions are not met.

Fourth, who has authority to define what counts as evidence? This question directs attention to the governance of epistemic infrastructure. It asks who participates in setting categories, designing metrics, curating data, and validating models. It recognises that epistemic filters embody power, and that legitimate systems require inclusive processes for determining what counts.

These questions can be applied at multiple stages: in dataset design, following Gebru and colleagues' (2021) call for rigorous documentation of datasets' origins, limitations, and intended uses; in model development, interrogating what patterns the model is learning and what it cannot learn; in deployment planning, assessing which decisions will be mediated by the model and with what opportunities for human judgment; in monitoring and evaluation, tracking not only performance metrics but also what escapes measurement.

What Success Looks Like

Short-term (1-3 years):
EDD questions integrated into procurement standards, dataset documentation, and model development practices. RFPs require epistemic exclusion analysis. Model cards include "structural inadmissibility" sections.

Medium-term (3-7 years):
Institutions demonstrate capacity to revise systems when blind spots become apparent. Affected communities have formal governance roles. Documented cases of system redesign based on community input. Measurable reduction in "lost to follow-up" rates for marginalised groups.

Long-term (7-20 years):
Epistemic diversity preserved across institutional domains. Democratic legitimacy sustained despite AI integration. No cascading epistemic failures. Sustained trust in institutions among groups historically burdened by legibility demands.

The Counterfactual

"What Happens if we ignore this"

Organised Epistemic Irresponsibility:

Exclusion occurs without any actor being accountable

Eroded Legitimacy

Populations whose knowledge is systematically excluded withdraw trust from institutions

Cascading Vulnerability

A blind spot in one domain propagates through interconnected systems

Correlated Blind Spots

Healthcare, education, welfare, and criminal justice systems inherit identical omissions

Lock-in

Epistemic filters embedded in foundation models become irreversible as switching costs mount

A brief example illustrates the approach. Consider using language models in healthcare triage or clinical decision support. Epistemic Due Diligence would ask: whose knowledge is excluded? Patients who do not speak dominant languages; patients whose symptom descriptions do not match typical presentations; patients with rare or under-documented conditions; patients whose embodied experience of illness resists textual description.

What likely lies outside the frame? Tacit knowledge of experienced clinicians; embodied knowledge of patients living with chronic conditions; oral traditions of community health practices; relational knowledge emerging from continuity of care. Which decisions should remain revisable? Diagnoses relying on pattern-matching from limited data; triage decisions that could cascade into delayed care; resource allocations that could entrench disparities. Who decides what counts as evidence? Currently, primarily those who design and deploy models—tech companies, healthcare administrators, researchers—with limited input from patients, communities, or frontline clinicians. Asking these questions does not automatically produce answers, but it shifts the burden of justification: it requires those building and deploying systems to account for what they are missing.

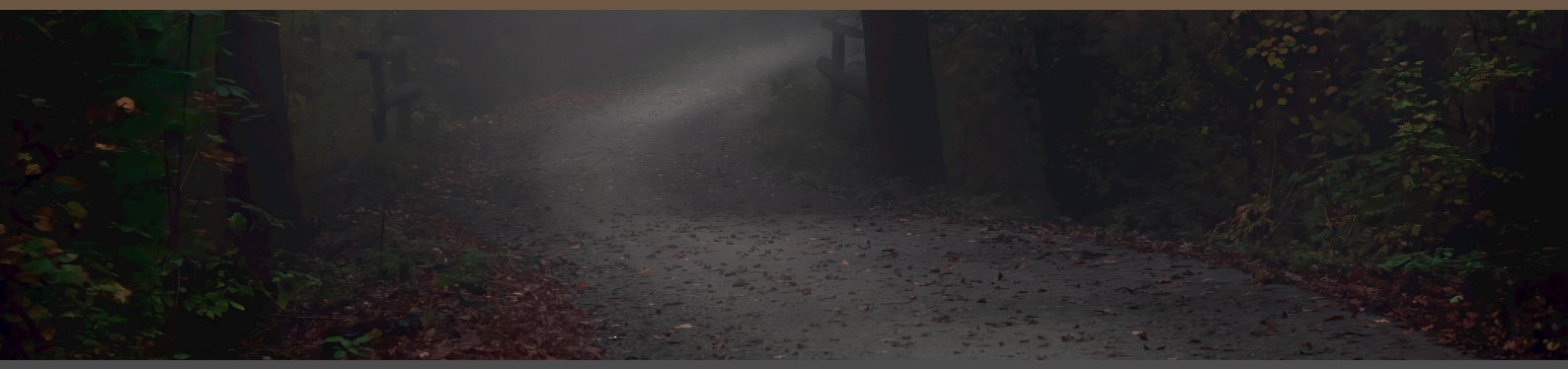
Epistemic Due Diligence could take institutional forms: independent review panels including affected communities; cross-disciplinary red teams tasked with identifying blind spots; data design reviews interrogating category choices before collection; participatory processes involving those who will bear the burden of legibility in shaping the systems that impose it. Such review should occur before procurement or deployment, involve affected communities and frontline practitioners, document known exclusions, and specify triggers for revision when exclusion becomes visible in practice. These forms would need adaptation to context, but the underlying principle is consistent: those who design epistemic infrastructure should be accountable for its filters.

A crucial limit must be acknowledged: Epistemic Due Diligence does not abolish filtering. It does not promise a system that captures all knowledge, includes all voices, or eliminates all bias. Such a system is impossible. What it offers instead is a practice of making filters visible, acknowledging what is being missed, and creating mechanisms for contestation and revision. It is due diligence not in the sense of eliminating risk but in the sense of knowing what risks are being run, and by whom. The goal is not perfect knowledge but accountable ignorance—ignorance that is recognised, named, and made revisable rather than ignorance denied, naturalised, and embedded in infrastructure.

The central claim of this paper is that the design of epistemic infrastructures should be evaluated not only for bias in outputs, but for the kinds of knowledge they exclude, burden, or render inadmissible from the outset.

Chapter 8

Conclusion



Chapter 8

Conclusion

Institutions act through epistemic filters. This is not a pathology but a condition of possibility: to govern at scale is to simplify, to categorise, to measure, and to represent. Yet these filters are not neutral. They embody judgments about what matters—decisions that become embedded in forms, databases, and routines, shaping countless further choices without requiring re-authorisation. The problem is not that institutions simplify, but that they forget they are simplifying, mistaking their maps for the territory and their metrics for the phenomena they purport to measure.

These filters impose uneven burdens of legibility. Some forms of knowledge pass easily because they are produced in recognised formats, by recognised actors, and in recognised languages. Others enter only through costly translation—the patient performing credibility, the community producing indicators, the marginalised group rendering its experience in terms it did not design. And some forms of knowledge never enter at all: the tacit, the embodied, the oral, the local, the spiritual, the relational—ways of knowing that remain structurally inadmissible because there are no institutional slots for them. The distribution of these burdens is not random but patterned by power. Those already marginalised bear the greatest costs of becoming legible, while those who cannot become legible face exclusion naturalised as absence, non-compliance, or lack of evidence.

Foundation models inherit these dynamics and amplify them. Trained on archives shaped by centuries of filtering, they learn from what survived to be written, digitised, and preserved. As they become epistemic infrastructure—mediating search, classification, summarisation, and decision-making across domains—they do not create new filters so much as consolidate and generalise existing ones. The result may be correlated blind spots and cascading vulnerabilities: when the same systems mediate healthcare, education, employment, and welfare, the same exclusions can propagate across institutional domains, compounding invisibility and eroding legitimacy.

Historically, the burden of legibility has fallen on the excluded. It is the excluded who must translate, perform credibility, and suppress aspects of their experience to become institutionally visible. A defensible institutional future would require institutions to shoulder more of that burden themselves—to build systems that reach toward what they cannot easily capture, to acknowledge what they are missing, and to create mechanisms through which those systematically unheard can contest the terms of their exclusion. This is not a call for epistemic perfectionism but for epistemic humility: the recognition that institutional knowledge is always partial, always filtered, and always shaped by decisions that could have been otherwise.

Epistemic Due Diligence is one attempt to operationalise that humility. By asking whose knowledge is excluded, what forms of understanding remain outside the frame, and who has authority to define what counts as evidence, institutions can begin to make their epistemic filters visible and contestable. The goal is not perfect knowledge, but accountable ignorance: ignorance that is recognised, examined, and made revisable rather than denied and embedded in infrastructure.

The planes that did not return carried information the visible data could not supply. The voices that never reach institutional awareness carry information too. The question is whether institutions can learn to look for them—and whether they can be made answerable to what they find.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), 100205.

Bowker, G. C., & Star, S. L. (1999). Sorting things out: Classification and its consequences. MIT Press.

Carel, H. (2016). Phenomenology of illness. Oxford University Press.

Cassam, Q. (2019). Vices of the mind: From the intellectual to the political. Oxford University Press.

Chambers, R. (1997). Whose reality counts? Putting the first last. Intermediate Technology Publications.

David, P. A. (1985). Clio and the economics of QWERTY. American Economic Review, 75(2), 332-337.

Dotson, K. (2012). A cautionary tale: On limiting epistemic oppression. Frontiers: A Journal of Women Studies, 33(1), 24-47.

Dotson, K. (2014). Conceptualizing epistemic oppression. Social Epistemology, 28(2), 115-138.

Eyben, R. (2013). Uncovering the politics of 'evidence' and 'results'. In R. Eyben, I. Guijt, C. Roche, & C. Shutt (Eds.), The politics of evidence and results in international development (pp. 19-38). Practical Action Publishing.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. American Political Science Review, 94(2), 251-267.

Pohlhaus, G. (2012). Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance. Hypatia, 27(4), 715-735.

Polanyi, M. (1966). The tacit dimension. Doubleday.

Porter, T. M. (1995). Trust in numbers: The pursuit of objectivity in science and public life. Princeton University Press.

Scott, J. C. (1998). Seeing like a state: How certain schemes to improve the human condition have failed. Yale University Press.

Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. Statistical Research Group, Columbia University. Reprinted in Journal of the American Statistical Association (1984), 79(386), 263-269.

Warren, M. E. (2017). A problem-based approach to democratic theory. American Political Science Review, 111(1), 39-53.

Werner, A., & Malterud, K. (2003). It is hard work behaving as a credible patient: Encounters between women with chronic pain and their doctors. Social Science & Medicine, 57(8), 1409-1419.

Zoglauer, T. (2022). Die Wahrheitskrise: Über Verschwörungstheorien, Fake News und die Filterblase. Reclam.

This page intentionally left blanc

After the Paper:
Recommendations & Research

The Core Argument

The white paper "Epistemic Due Diligence" accomplishes something rare: it identifies a genuinely novel research domain at the intersection of three established fields—philosophy of epistemology, science and technology studies (STS), and AI governance—and provides both a diagnostic framework and a roadmap of unanswered questions. What the paper does not do, and could not do in twelve weeks, is develop the methodological toolkit, empirical grounding, and institutional design work required to translate EDD from hypothesis into practice.

That translation is the work of a multi-year research program. And it is urgent.

Why This Research Program Matters Now

The timing is critical. Foundation models are currently being integrated into institutional infrastructure at unprecedented speed. The epistemic filters they inherit and amplify are hardening into place. Decisions about what counts as knowledge—who is heard, what is seen, what is missed—are being made today by engineers and product managers, not by philosophers or affected communities. Within five years, many of these filters will be locked in, their contingency forgotten, their exclusions naturalised.

Research that intervenes upstream, before lock-in occurs, has the potential for extraordinary impact. This is a narrow window.

The Research Agenda: Four Pillars

The questions arising section of the paper can be organised into four coherent research streams, each of which could sustain a doctoral thesis or a multi-year postdoctoral project. Together, they constitute a research program.

Pillar One: The Detection Problem

Core Question: How can structural inadmissibility be detected when it leaves no institutional trace?

This is the foundational methodological challenge. The paper argues that some forms of knowledge are not merely difficult to capture but impossible to capture without distortion. But if such knowledge leaves no trace, how can any due diligence process register its absence? This is not a problem that philosophy alone can solve. It requires empirical investigation of how exclusion manifests indirectly—through patterns of system failure, through rates of appeal, through ethnographic observation of frontline practice, through the testimony of those who bear the burden of legibility.

Research Questions:

- Can near-misses and edge cases serve as signals of epistemic blind spots? When a system fails in ways that surprise its designers, does that failure often trace to knowledge that was structurally excluded from training data?

- What methods from science and technology studies—ethnography, participant observation, qualitative interviewing—can be adapted to detect epistemic exclusion in operational systems?

- Can computational methods surface what a model cannot represent? Are there adversarial approaches that reveal epistemic boundaries rather than performance failures?

Contribution:

This stream would develop the methodological toolkit for EDD, moving it from diagnostic questions to empirically grounded detection practices.

Pillar Two: The Preservation Paradox

Core Question: What is lost when the illegible is made legible, and should some knowledge remain structurally inadmissible?

The paper identifies a tension that is both theoretically profound and practically urgent. EDD seeks to make institutions accountable to excluded knowledge, but the very act of rendering such knowledge institutionally visible may transform or damage it. When oral traditions are written down, when embodied practices are codified, when relational goods are measured—something is lost. But what? And is that loss sometimes acceptable, and sometimes not?

This is not a question that can be answered abstractly. It requires detailed case studies of knowledge forms that have undergone, or resisted, translation into institutional categories.

Research Questions:

- What can be learned from historical cases of knowledge translation—the codification of customary law, the digitisation of oral histories, the standardisation of traditional medicine? What was lost, what was gained, and who benefited?

- Are there knowledge forms that communities wish to protect from institutional capture? What would epistemic sovereignty look like in practice, and how could EDD respect it?

- How should institutions distinguish between valuable knowledge that is structurally excluded and knowledge that is rightly excluded because it cannot withstand scrutiny? Is there a workable normative framework for making this distinction?

Contribution:

This stream would develop the normative boundaries of EDD, establishing when epistemic inclusion is appropriate and when it constitutes a form of extraction.

Pillar Three: The Governance Gap

Core Question: What institutional forms can sustain epistemic accountability over time, and how should epistemic disagreement be resolved?

The paper's fourth diagnostic question—who has authority to define what counts as evidence—opens onto a landscape of unresolved governance challenges. If epistemic filters embody judgments about what matters, and if those judgments are properly contestable, then institutions need mechanisms for contestation, deliberation, and resolution. But existing governance frameworks for AI—focused on transparency, fairness, and non-discrimination—are poorly equipped to handle disputes about what counts as knowledge in the first place.

This stream would design and test institutional forms for epistemic governance.

Research Questions:

- What would epistemic review panels look like? Who should sit on them, what authority should they have, and how should they deliberate?

- How should participatory processes be designed to avoid co-optation? When communities are invited to participate in system design, how can they be given real authority rather than merely providing epistemic labour to legitimise predetermined outcomes?

- What legal frameworks could assign liability for epistemic harm? If organised epistemic irresponsibility produces systematic exclusion, who should be accountable, and through what mechanisms?

- How should institutions handle irreconcilable epistemic disagreement? When communities, practitioners, and technical experts disagree about what knowledge matters, what procedures should govern resolution?

Contribution:

This stream would move from critique to institutional design, producing concrete proposals for governance structures that can operationalise epistemic accountability.

Pillar Four: Empirical Mapping of Epistemic Narrowing

Core Question: Is epistemic narrowing accelerating, and can it be measured?

The paper hypothesises that foundation models consolidate and amplify epistemic filters, creating correlated blind spots across institutional domains. But this hypothesis remains untested. Testing it requires empirical methods for measuring epistemic diversity—or its absence—in knowledge systems over time.

This stream would develop metrics and conduct large-scale empirical studies to establish the baseline conditions against which interventions can be evaluated.

Research Questions:

- Can epistemic diversity be measured? What indicators would capture the range of knowledge forms present in a given domain, and the distribution of epistemic authority across groups?

- Is there evidence of epistemic narrowing in domains where foundation models have been deployed? For example, has the use of AI in hiring reduced the range of qualifications considered? Has its use in healthcare narrowed the forms of patient experience that register as relevant`?

- What would epistemic resilience look like? Are there systems or domains that have successfully resisted narrowing, and what can be learned from them?

Contribution:

This stream would provide the empirical foundation for the entire EDD framework, establishing whether the problem it diagnoses is real and measurable, and providing baselines against which interventions can be assessed.

In Closing

No single researcher could pursue all four pillars simultaneously, but the pillars are interdependent. Methodological work on detection (Pillar One) informs empirical measurement (Pillar Four). Normative work on the preservation paradox (Pillar Two) constrains institutional design (Pillar Three). Empirical findings (Pillar Four) reveal which governance gaps are most urgent (Pillar Three).

A doctoral thesis could plausibly address one pillar in depth while contributing to the others. A postdoctoral fellowship could extend the work across pillars. A research group or centre could pursue the full program. This white paper has done the essential preliminary work: it has defined the problem, synthesised the relevant literatures, and generated a coherent set of research questions.

What it has not done—and could not do in twelve weeks—is the sustained empirical, methodological, and institutional design work required to translate EDD from hypothesis into practice.

That is the work of a much larger project.