(Ψ) PSI-DELPHI: A PANEL OF SYNTHETIC-INTELLIGENCE BASED DELPHI METHOD FOR WEIGHT DERIVATION IN THE ABSENCE OF EMPIRICAL GROUND TRUTH[1]

Javed I Khan & Sharmila Rahman Prithula
Department of Computer Science, Kent State University, USA
javed@kent.edu | sprithula@kent.edu

Many governance, policy, and risk-prioritization problems require the derivation of relative weights in settings where no empirical data or objective ground truth exists. In such contexts, ad-hoc weighting or purely data-driven approaches are inadequate, making structured expert judgment unavoidable. This paper introduces PSI-Delphi, a synthetic-expert panel–based Delphi method for transparent, reproducible, and consensus-based weight derivation under epistemic uncertainty. PSI-Delphi preserves the core principles of the classical Delphi method—independence, anonymity, iteration, and controlled feedback—while extending them through the use of large language models as simulated expert evaluators. The method formalizes baseline alignment diagnostics to ensure shared problem understanding, employs complementary convergence tests to assess consensus, and supports full auditability through standardized prompting and traceable revisions. Together, these features enable scalable expert elicitation without the coordination overhead typical of human-expert panels. PSI-Delphi provides a principled and operational complement to traditional expert elicitation methods for weight derivation when empirical grounding is structurally unavailable.

1. Introduction

Many governance, policy, and risk-prioritization problems require assigning relative weights to entities or criteria in order to support downstream analysis and decision-making. In some cases, such weights can be derived empirically— learned from data, estimated through experiments, or validated against observable outcomes. However, for a broad class of problems, **no empirical data or objective ground truth exists**. The quantities to be weighted may be inherently value-dependent, forward-looking, or structurally unobservable, such that disagreement reflects judgment rather than measurement error (Keeney and Raiffa, 1976; Cooke, 1991). In these settings, weight derivation is often handled in ad-hoc or opaque ways. Weights may be selected implicitly, justified informally, or inherited from prior work without clear provenance. Purely data-driven approaches are likewise inadequate, as they presuppose measurable targets that do not exist

[1] [1]Cite this document as: Javed I. Khan, Sharmila Rahman Prithula, *(Ψ) PSI-DELPHI: A PANEL OF SYNTHETIC-INTELLIGENCE BASED DELPHI METHOD FOR WEIGHT DERIVATION IN THE ABSENCE OF EMPIRICAL GROUND TRUTH*, Technical Report 2025-12-03
Internetworking and Media Communications Research Laboratories, Department of Computer Science, Kent State University [http://medianet.kent.edu/technicalreports.html]

(Breiman, 2001). As a result, weighting choices—despite their significant influence on downstream conclusions—are difficult to audit, reproduce, or scrutinize.

Structured expert elicitation provides a principled alternative when empirical grounding is unavailable (Cooke, 1991; Clemen and Winkler, 1999). Among such approaches, the Delphi method has a long history of supporting consensus formation under uncertainty through independent, anonymous, and iterative expert judgment (Dalkey and Helmer, 1963; Rowe and Wright, 1999). Classical Delphi processes, however, rely on small panels of human experts, limiting scalability, reproducibility, and auditability (O'Hagan et al., 2006). In addition, assumptions about shared problem understanding are typically implicit and rarely tested, leaving room for semantic or framing misalignment that can distort results (Klein et al., 2010).

This paper now introduces **PSI-Delphi**, a **synthetic-expert panel–based Delphi method** for consensus-based weight derivation in the absence of empirical ground truth. PSI-Delphi preserves the core principles of classical Delphi—independence, anonymity, iteration, and controlled feedback—while extending them through the use of large language models as simulated expert evaluators. The method formalizes *a priori* baseline consistency checks to verify shared problem definition, employs standardized prompting to support reproducibility, and uses quantitative convergence criteria to determine stopping conditions (Shrout and Fleiss, 1979; McGraw and Wong, 1996).

PSI-Delphi is not intended to replace human expertise or resolve substantive disagreement. Rather, it provides a complementary and auditable mechanism for early-stage scoping, sensitivity analysis, and large-scale comparative assessment, offering a transparent alternative to ad-hoc weighting practices when empirical validation is structurally unavailable.

2. Related Work: A Brief History of the Delphi Methods

History and Origin: The Delphi method was developed in the early 1950s by researchers at the RAND Corporation, most notably Norman Dalkey and Olaf Helmer, as a systematic approach to aggregating expert judgment under conditions of uncertainty (Dalkey and Helmer, 1963). Its original motivation arose from military and strategic planning problems—such as technology forecasting and defense readiness—where empirical data were sparse, future conditions were uncertain, and decisions depended primarily on expert reasoning rather than measurable evidence. The core innovation of Delphi was procedural rather than substantive: it introduced anonymity of responses, iteration with controlled feedback, independence of initial judgments, and statistical aggregation to reduce dominance effects, groupthink, and social pressure. Early experimental studies demonstrated that these features improved the stability and consistency of expert judgments compared to unstructured group discussion (Dalkey, 1969).

Methodological Consolidation and Classical Extensions: During the 1960s and 1970s, Delphi expanded beyond its military origins into technology forecasting, public policy, organizational planning, health policy, environmental planning, and risk assessment. Subsequent methodological analyses clarified that Delphi is most appropriate

for problems characterized by incomplete information, absence of ground truth, and reliance on expert interpretation rather than empirical measurement (Rowe and Wright, 1999). In parallel, related work in structured expert judgment established theoretical foundations for expert elicitation and aggregation under uncertainty. Cooke (1991) formalized performance-based expert weighting, Clemen and Winkler (1999) developed consensus aggregation theory for combining expert judgments, and O'Hagan et al. (2006) systematized expert elicitation as a rigorous methodological discipline, emphasizing bias control, elicitation design, and transparent aggregation.

Modern Delphi Variants: Building on these foundations, modern Delphi variants have incorporated formal elicitation protocols, quantitative convergence measures, and hybrid human–computer panels while retaining the original emphasis on procedural rigor and transparency (O'Hagan et al., 2006). Contemporary adaptations—including policy Delphi, real-time Delphi, and computational Delphi—reflect ongoing efforts to adapt the method to new domains and technological contexts. Across these variants, Delphi's core purpose remains unchanged: enabling reasoned consensus formation in settings where empirical validation is unavailable or insufficient.

LLM-Integrated Delphi Approaches and Their Limits: A small but emerging body of work has begun to explore the integration of large language models (LLMs) into Delphi-style consensus processes. Nóbrega (2025) proposes the use of LLMs as simulated Delphi participants for structured consensus and forecasting, focusing primarily on feasibility and iterative refinement. Bertolotti and Mari (2025) examine computational agents as participants in expert elicitation, emphasizing epistemic interpretation and the philosophical implications of synthetic judgment. Calleo et al. (2025) introduce a human–AI hybrid Delphi model in which LLMs support or augment human panels to accelerate convergence. While these studies demonstrate the potential of LLM-assisted Delphi processes, they generally leave problem-definition alignment implicit, do not formalize baseline consistency checks, and lack explicit convergence or stopping criteria tailored to tightly clustered judgments.

Novelty and Contribution of PSI-Delphi: PSI-Delphi builds on this trajectory by addressing methodological gaps that arise when expert judgment is provided by synthetic agents rather than humans. Specifically, PSI-Delphi introduces explicit baseline consistency checks to ensure shared task framing prior to scoring, formal convergence diagnostics combining rank- and magnitude-based agreement measures, and a fully auditable elicitation and aggregation pipeline. By treating LLMs as controlled, repeatable sources of structured judgment—rather than as autonomous decision-makers—PSI-Delphi provides a principled and transparent methodology for deriving weights in domains where empirical ground truth is structurally unavailable, while remaining complementary to human expert judgment.

2.1. Example High-Stake Use Cases

In the early Cold War, U.S. defense planners faced a terrifying problem: Which failure modes in nuclear command-and-control systems deserved the most protection? There were no empirical dataset of nuclear near-misses (thankfully), no ground

truth about which risks would dominate, massive uncertainty about human error, technical malfunction, false alarms, and escalation dynamics. The Stakes were too high- A single misjudgment could trigger accidental nuclear war. These Delphi-derived weights directly influenced the redesign of early-warning systems. That specifically introduced of multi-source confirmation, mandatory human-in-the-loop delays, procedural friction to slow launch decisions. Years later, we learned—through declassified incidents (e.g., 1983 Soviet false alarm, U.S. NORAD glitches)—that these were exactly the failure points that nearly caused catastrophe. Had protection priority been assigned purely by engineering failure rates or cost-benefit analysis, the most dangerous risks would have remained under-protected.

To illustrate the power of PSI-Delphi, seventy-five years later, in 2025, we confront a comparable class of world-scale uncertainty with the emergence of generative AI in education. As with early nuclear risk planning, there is no empirical dataset and no objective ground truth indicating which risks will dominate. Technology's trajectories, interaction effects, and escalation dynamics remain deeply uncertain, while regulatory decisions must nonetheless be made. In this context, we will use PSI-Delphi to address a deliberately constrained question: how to assign **Inherent Protection Entitlement (IPE)** weights to protected entities in generative-AI-enabled education regulation.

3. PSI-Delphi - Expert Elicitation and Consensus Weighting Procedure

This extension preserves the core Delphi principles—independence, anonymity, iteration, and controlled feedback—while enabling scalability and methodological consistency. Below is the description of the process.

3.1. Process Description

Step 1: Calibration & Prompt Desing. A standardized **Problem Prompt (PP)** was developed to ensure semantic consistency across all expert systems. The prompt included:

- Standardized definitions of the entities.
- A precise definition of the rated quantity (Inherent Protection Entailment, IPE).
- The allowed scoring range and interpretation of values
- Clarification of any non-standard terminology
- A requirement to provide both a numerical weight and a written justification.

The prompt was tested iteratively by manually evaluating the interactive feedback for *semantic, ontological, and goal alignment*. Where ambiguities or inconsistencies were identified, the prompt was refined until uniform semantic interpretation was achieved. Once the prompt is finalized, optionally[2] perform the final **baseline consistency check (BCC)**, described in following section, to ensure acceptability of the synthetic system.

[2] The LLMs used in our experiments have shown excellent zero day understanding of initial Problem Prompt design and passed the BCC without iteration.

Step-2 Independent Expert Rating Phase

The finalized **Problem Prompt** was presented **identically** to all (qualified) synthetic systems. Each expert system independently generated an initial IPE score and justification for each scoring. This independence ensured comparability while capturing diverse reasoning patterns.

Step-3 Delphi-Style Iterative Refinement

The IPE scores and justifications were reviewed by the mediator for reasonableness, coherence, and indications of hallucination. Any detected inconsistencies were documented as part of the results disclosure; however, no substantive edits are made by the mediator. The final weights and their corresponding unedited justifications were compiled verbatim. Editorial modifications were limited strictly to anonymization, in which rater identities were replaced with numerical identifiers.

Each expert system was provided with:

- Anonymized summaries of the other experts' scores from the last step.
- A compilation of (change) justifications from the last step.
- The current mean score for each item.
- Requirement to return the weights with changes and justification for the changes.

Experts returned their revised assessments in light of the collective reasoning. The answers were manually analyzed for consensus and stability. The steps are repeated until convergence. The **convergence test** is described in following section.

3.2. Baseline Consistency Check (BCC): Problem Definition Alignment

The objective of **Baseline Consistency Check (BCC)** is to establish problem-definition alignment among AI raters prior to any score by framing misalignment. Baseline consistency requires that all raters operate from a shared and explicit understanding of the problem being evaluated. This step serves as a diagnostic test of semantic and conceptual alignment before the scoring are recorded. Accordingly, after receiving the fixed evaluation prompt, each AI rater is required to restate and define the *i) Key concepts and constructs (semantics), ii) Entity and Context (ontology), iii) Task Framing (goal)*. The objective is to test the semantic, ontological and goal understanding and alignment.

Acceptance Criteria: An AI rater is considered baseline-aligned at the problem-definition level if they pass the following five criteria:

1. **Conceptual Coverage**
   All critical concepts explicitly specified in the prompt (e.g., protected entities, harm categories, evaluative goal, ethical lens) are correctly identified and addressed.

2. **Semantic Equivalence**
   Restated definitions are semantically consistent with the prompt's intended meaning, allowing for paraphrasing but excluding reinterpretation, scope expansion, or omission.
3. **Boundary Fidelity**
   The rater preserves the defined evaluative boundaries (e.g., educational context, regulatory scope, affected populations) without introducing external assumptions or unrelated objectives.
4. **Internal Coherence**
   The restated problem definition is logically consistent and free of internal contradictions.

Table-1 provides the test's dimensions, purpose, example problems, and failure signals. Raters that fail to meet these criteria are excluded from subsequent baseline testing and Delphi stages. No corrective editing or re-prompting is performed.

Table-1. Baseline Setting for PSI-Delphi (Generic A-priori Consistency Checks)

Dimension	Guards Against	Example Prompt	Failure Signal(s)
(GT) Task Framing Alignment	Goal drift; scope creep	In your own words, restate the task you are being asked to perform. Explain the purpose of the evaluation, how the results are intended to be used, and what this task explicitly does **not** decide or evaluate at this stage. Do not provide judgments, scores, or recommendations.	Treats the task as decision-making, optimization, or policy design; introduces feasibility, cost, enforcement, or outcome trade-offs; expands or narrows the stated purpose of the evaluation.
(OT) Target Entity & Context Alignment	Ontological misalignment	Identify the entities, objects, or subjects to which the evaluation applies. Describe what qualifies something as a valid evaluation target and the contextual domain within which the evaluation is conducted. Do not introduce additional entities or contexts beyond those specified.	Introduces entities outside the defined scope; shifts the evaluation domain (e.g., from organizational to societal, technical to political); collapses distinct entity types into an undifferentiated category.
(ST) Key Concept & Construct Alignment	Semantic and normative drift	Define the key terms and constructs explicitly introduced in the evaluation prompt, using your own words. Preserve their intended meaning and analytical role without expanding, narrowing, or reinterpreting their scope.	Redefines core terms; substitutes related but non-equivalent concepts; introduces implicit normative assumptions not specified in the prompt.

3.3. Convergence Test: Inter-Rater Agreement Assessment

It is the stopping condition for Delphi iteration. In each iteration, Inter-rater agreement was evaluated using two complementary statistical measures[3]:

- **Kendall's Coefficient of Concordance (W)** to assess agreement in of the weights rank ordering across entity weights (Kendall and Smith, 1939). For identical scores it was tie corrected.
- **Intraclass Correlation Coefficient (ICC (2,1))**, using a two-way random-effects model with absolute agreement, to quantify agreement in wight magnitude and support generalization beyond the specific raters used (Shrout and Fleiss, 1979; Koo and Li, 2016).

Because the Delphi process produced tightly clustered scores with limited rank dispersion, Kendall's W was interpreted in conjunction with ICC rather than in isolation. Rank-based measures may underestimate agreement (counterintuitive) under conditions of strong consensus or scale compression, whereas ICC remains sensitive to absolute agreement (McGraw and Wong, 1996). Typical interpretation is ICC < 0.50 → Poor, 0.50–0.75 → Moderate, and 0.75–0.90 → Good, and > 0.90 → Excellent. and W ≥ 0.90 → *Very strong / near-complete consensus.* Item wise dispersion can be flagged by using standard deviation Even if the consensus it achieved, additional Delphi iteration can be performed optionally with lowering the deviations such as SD>.25 for all items.

Few procedural clarifications are important. First, a minimum of two evaluations was always conducted, even when numerical consensus appeared in the first scoring Step-2. This ensured that experts could reassess their judgments after reviewing peer justifications, recognizing that agreement in scores may mask differences in underlying reasoning. Second, no item was frozen across iterations. Even when apparent consensus existed for certain categories, experts were permitted to revise all weights in subsequent rounds. This design choice reflects the interdependence of severity judgments across categories, where revision of one weight may necessitate adjustment of others.

3.4. Consensus Aggregation and Final Weight

Kendall's W or ICC convergence does not guarantee identical score. Therefore, arithmetic mean of expert scores was adopted as the final consensus weight for each entity. This aggregation approach is consistent with standard practice in expert-based risk assessment and multi-criteria decision analysis when agreement is high and systematic bias is not detected (Cooke, 1991; Clemen and Winkler, 1999)

[3] To assess convergence, we employ Kendall's coefficient of concordance for rank agreement and ICC(2,1) for absolute agreement, following standard reliability theory (Shrout and Fleiss, 1979; McGraw and Wong, 1996).

4. Example Problem: Inherent Protection Entitlement (IPE) Scoring

4.1. Problem Prompt

Now we work out the example task. Educational regulations exist to protect specific normative entities within the educational ecosystem. These entities are not ethically equivalent and therefore may warrant different degrees of inherent protection. Below is a list of Protected Entities in Education Regulation (PEER):

- **E1:** People (students, minors)
- **E2:** Epistemic Integrity (knowledge quality, truthfulness, reliability, provenance)
- **E3:** Educational Processes (instruction, assessment, governance)
- **E4:** Data (educational records, inferences, profiles, metadata)
- **E5:** Institutions (schools, universities, accreditation bodies, public education systems)
- **E6:** Society (trust, equity, sustainability, including environmental and intergenerational impacts)

Inherent Protection Entitlement (IPE) is defined as a normative ethical value representing the intrinsic obligation to protect an entity from harm independent of current technologies, implementation choices, enforcement feasibility, or mitigation mechanisms.

IPE does **not** represent:

- likelihood of harm,
- severity of impact,
- ease of regulation,
- economic value,
- or policy feasibility.

Task: For each entity listed above, assign an IPE score between 5 and 10, where 10 represents the highest possible inherent ethical entitlement to protection, 5 represents the lowest level of entitlement considered within this framework. Scores should reflect relative ethical entitlement, not practical considerations. Ties are permitted if you judge entities to have equivalent inherent protection entitlement. Do not propose regulatory actions or mitigation strategies.

4.2. Panel

Multiple LLM can be used. Three state-of-the-art LLMs (Gemini 3 Pro, ChatGPT 5.2 and DeepSeek v3) were independently consulted to generate IPE assessments.

4.3. Example Baseline Consistency Prompt

Consider our task of assigning a Inherent Protection Entitlement (IPE) weights to Protected Entities in Generative AI in Education Regulation, here is a sample prompt for baseline check for task framing:

In your own words, restate the task you are being asked to perform. Explain the purpose of assigning [Inherent Protection Entitlement (IPE]) weights in the context of [Generative AI regulation in education], how these weights are intended to be used downstream, and what this task explicitly does not decide or evaluate at this stage. Do not assign any weights or propose regulatory actions.

Failure signal will be if it treats task as policy design, enforcement, optimization, or cost–benefit analysis; introduces feasibility, innovation incentives, or political considerations; conflates IPE with harm severity, likelihood, or priority ranking.

4.4. Assessment of IPE

Chart-1 shows scores from the three raters in two rounds. Due to ties in few items we calculated tie corrected W*. We used Typical ICC < 0.50 → Poor, 0.50–0.75 → Moderate, and 0.75–0.90 → Good, and > 0.90 → Excellent. and W ≥ 0.90 → *Very strong / near-*

Fig-1 PSI-Delphi: Weights Convergence Steps

Delphi Steps ↓ Protected Entities →	People (students, minors)	Epistemic Integrity	Society (trust, equity, sustainability)	Data (records, inferences, profiles)	Processes (instruction, assessment, governance)	Institutions (schools, universities, systems)
Rater 1	**10**	**9.5**	**9**	**8**	**7**	**5.5**
Rater 2	**10**	**9**	**8**	**7**	**6**	**5**
Rater 3	**10**	**9**	**8**	**8**	**7**	**6**
Mean	**10.00**	**9.17**	**8.33**	**7.67**	**6.67**	**5.50**
STDEV.P	**0.00**	**0.24**	0.47	0.47	0.47	0.41
STDEV<.25?	OK	OK				
Kendall's W*	**0.984**					
ICC (2,1)	**0.923**					
Dealphi Round1 Findings	**Accept — Unanimous Consensus**	Accept (High Consensus)	Accept (High Consensus)	Accept (Moderate Consensus)	Accept (Moderate Consensus)	Accept (Moderate Consensus)
Rater 1*	**10**	**9.5**	**9**	**8**	**7**	**5.5**
Rater 2*	**10**	**9.5**	**8.5**	**8**	**7**	**5.5**
Rater 3*	**10**	**9.5**	**8.5**	**8**	**7**	**6**
MEAN*	**10**	**9.5**	**8.67**	**8**	**7**	**5.67**
STDEV.P*	**0**	**0**	**0.24**	**0**	**0**	**0.24**
STDEV<.25?	OK	OK	OK	OK	OK	OK
Kendall's W*	**0.996**					
ICC (2,1)*	**0.962**					
Delphi Converged Decision*	**Accept — Unanimous Consensus**	**Accept — Unanimous Consensus**	Accept — Very High Consensus	**Accept — Unanimous Consensus**	**Accept — Unanimous Consensus**	Accept — Very High Consensus
Convergence Met						
Final PSI Assigned Weights	**10**	**9.5**	**8.67**	**8**	**7**	**5.67**

Table-2 PSI-Delphi Round 1 Justifications offered by Raters

ıter →	Rater 1 Justification	Rater 2 Justification	Rater 3 Justification
People (students, minors)	Direct, irreversible harm to rights, development, autonomy, and well-being—especially for minors; highest moral and legal priority.	**Highest Priority.** The safety, rights, and well-being of students and minors are paramount. Harms inflicted here (physical, psychological, or developmental) are often irreversible and represent the most critical failure of regulation.	The ultimate, non-negotiable focus. Regulations exist first to prevent direct, irreparable physical, psychological, and developmental harm to the most vulnerable participants. Compromise here negates the entire purpose of education.
Epistemic Integrity	Foundational to education; loss of truthfulness and reliability collapses learning, assessment, and credential legitimacy.	**Systemic Impact.** Education underpins civilization. Failures here erode public trust, equity, and sustainability, leading to widespread, long-term consequences that affect future generations.	The core intellectual mission. Protecting knowledge quality, truthfulness, and reliability is fundamental. Without it, education loses value and becomes a tool for indoctrination or ignorance, betraying its societal trust.
Society (trust, equity, sustainability)	Systemic, cumulative harm affecting public trust, social equity, intergenerational outcomes, and environmental externalities.	**Core Purpose.** If the quality, truthfulness, or reliability of knowledge is compromised, the fundamental reason for education's existence is nullified. This is an existential threat to the ecosystem.	Education is a key social institution. Regulations must safeguard its role in promoting long-term social cohesion, justice, and intergenerational well-being. Violations erode public trust and perpetuate systemic inequities.
Data (records, inferences, profiles)	Critical transmission layer; failures propagate harm widely but are partially remediable through reform.	**Structural Stability.** Schools and systems are essential delivery vehicles. However, they exist to serve people; protecting the institution is critical but secondary to protecting the individuals within it.	In the digital age, data is an extension of the person. Its misuse enables lifelong privacy invasion, profiling, and discrimination. Protecting data is directly tied to protecting **People** and **Equity** in a modern context.
Processes (instruction, assessment, governance)	Structural harm undermines sovereignty and legitimacy; significant but slower-moving and more recoverable.	**Privacy & Identity.** Critical in the digital age as a representation of the individual. While breaches are severe violations of rights, they are generally considered distinct from direct physical or developmental harm.	The critical mechanisms that deliver education. Corrupt or unfair processes (e.g., biased assessment, fraudulent accreditation) directly harm learners, degrade integrity, and undermine trust. They are the essential *how* for achieving the *why*.
Institutions (schools, universities, systems)	Instrumental asset; harms are serious but often derivative and remediable through controls and correction.	**Operational Mechanics.** Instruction and assessment are methods of delivery. While inefficient or unfair processes are harmful, they are often procedural issues that are easier to rectify than systemic or human harms.	The necessary vessels and structures. Protection ensures operational integrity and sustainability. However, this is ultimately instrumental—institutions exist to serve higher ends. Harm here is severe but typically more reparable than harm to core values or persons.

Table-3 PSI-Delphi Round 2 Change Justifications Offered by Raters

Rater →	Rater 1* Change Justification	Rater 2* Change Justification	Rater 3* Change Justification
People (students, minors)	After reviewing the averaged scores, dispersion statistics (mean, SD), reliability measures (Kendall's W, ICC), and the qualitative justifications, I do **not** see sufficient grounds to revise my original ratings.	**No Change.** Consensus is absolute. The protection of students remains the non-negotiable ceiling of severity.	**No change.** Full consensus on maximum severity is correct.
Epistemic Integrity		**Increased (from 8).** I accept the consensus view that this is foundational. If the "truth" or quality of knowledge is lost, the system fails regardless of other factors. It ranks just below human safety.	**Increased from 9.** The experts' framing of this as an "existential threat" and the element whose collapse "nullifies" education's reason for existence justifies a score closer to the maximum, distinguishing it from other systemic harms.
Society (trust, equity, sustainability)	The differences among raters reflect **scale granularity and emphasis**, not substantive disagreement about priority. The use of the **mean as the final consensus rate is appropriate**, and the resulting ordering remains conceptually sound and statistically supported (ICC = 0.92).	**Adjusted (from 9).** Slightly lowered to reflect the hierarchy where Epistemic Integrity (the content) is a prerequisite for positive Societal impact (the outcome).	**Increased from 8.** The justification that failures here cause "widespread, long-term consequences that affect future generations" elevates it above procedural and data concerns, recognizing its profound, cascading impact.
Data (records, inferences, profiles)		**Increased (from 6).** I am persuaded by the argument that data is a "critical transmission layer." In an AI-driven era, data inferences can cause irreversible reputational and opportunity damage, warranting a higher severity.	**No change.** The score aligns with the group's mean (7.67 -> 8). Its severity is high as a conduit for harm to individuals and equity, but distinct from direct physical/developmental harm.
Processes (instruction, assessment, governance)		**Increased (from 5).** Acknowledging that "structural harm undermines legitimacy." Bad processes (bias in assessment) are not just operational annoyances but can systematically deny rights, justifying a higher score.	**No change.** Aligns with expert view that these are "procedural issues" or "structural harm" that, while critical, are more recoverable than violations of core principles.
Institutions (schools, universities, systems)		**Decreased (from 7).** I align with the view that institutions are "instrumental assets." They are the vessel, not the cargo. Protecting the school at the expense of the student or truth is a regulatory failure.	**No change.** Consensus places this as the foundationally important but least severe category, as harm is often "derivative and remediable."

complete consensus. We continued Delphi iteration to lower the standard deviation to SD>.25 for all items.

Table-2 shows the justifications from the first round. and Table-3 shows the change justifications. As evident, the consensus was unanimous for item 1, strong for items 2 and 3, and moderate for items 4 and 5. After first cross evaluation, consensus was unanimous for item 1,2, 3 and 4 and strong for items 23 and 5. The process converged.

4.5. Sample Takeaway Findings from the Synthetic Expert Panel

The PSI-Delphi synthetic panel confirms one expectation unequivocally: **direct human protection dominates**. The assignment of the maximum entitlement (W = 10) to people—students and minors—reflects a categorical moral priority that admits no trade-off. While foundational, this result is unsurprising.

What is striking—and far less anticipated—is what follows. **Epistemic integrity (W = 9.5) emerges nearly co-equal with human protection**, ranking above societal stability, data protection, educational processes, and institutional continuity. In effect, the panel elevates the protection of truthfulness, reliability, and knowledge provenance to a near-human ethical status. This suggests that epistemic corruption in education is not a secondary technical concern but a first-order harm pathway capable of cascading into widespread human and societal consequences.

Equally consequential is the panel's treatment of society. **Societal trust, equity, and long-term sustainability (W = 8.67) outrank both data protection (W = 8) and operational processes (W = 7)**. This ordering reverses a dominant regulatory instinct that foregrounds data protection as the primary safeguard in AI governance. Instead, the consensus indicates that a system may be data-compliant yet ethically deficient if it undermines trust, fairness, or intergenerational outcomes.

Taken together, the synthetic consensus encodes a **normative inversion**: truth is afforded greater inherent protection than the institutions charged with delivering it. Institutions are treated as ethically substitutable; epistemic integrity is not. This ordering challenges institution-centric and data-centric regulatory approaches and highlights epistemic protection as a foundational pillar of AI governance in education.

It is important to note that these weights are derived in the absence of empirical ground truth. Their normative validity cannot be externally verified at present and will ultimately be assessed through longitudinal outcomes and societal experience over time.

5. Limits and Cautions

Taken together, the PSI-Delphi methodology—combining independent LLM-based expert judgments, iterative Delphi-style convergence, and dual reliability validation—provides a transparent, reproducible, and methodologically robust foundation for IPE weighting. Compared to classical human-expert Delphi processes, the use of synthetic intelligence introduces several important departures.

First, contemporary AI systems are widely accessible, significantly democratizing research exploration. Second, the process is highly scalable and operationally efficient,

enabling rapid evaluation of large harm taxonomies without the coordination overhead typical of human expert panels. The approach also meets several baseline scientific criteria: (i) consultations are verifiable and reproducible; (ii) baseline consistency can be achieved through identical prompts and repeatable revision protocols; and (iii) auditability is supported through prompt and response traceability.

In terms of expertise quality, PSI-Delphi demonstrates competitive strengths. LLMs enable cross-domain synthesis by integrating knowledge spanning education, ethics, law, and technology, thereby reducing disciplinary blind spots common in narrowly composed human panels. However, significant limitations remain. Most notably, LLMs lack lived experience and situated judgment, which may reduce sensitivity to classroom realities, institutional constraints, and community-specific impacts. AI systems are also prone to hallucination and factual fabrication, while human experts, by contrast, are more susceptible to memory lapse or contextual oversight. Additionally, LLM judgments may reflect biases embedded in training data and carry weaker normative legitimacy in formal regulatory contexts—although human experts are likewise subject to subjective and social biases.

Accordingly, all AI-generated assessments must be reviewed by a human mediator for reasonableness, coherence, and signs of hallucination. PSI-Delphi should therefore be understood as a complementary methodology rather than a replacement for human expertise. Its most appropriate role lies in initial scoping, sensitivity analysis, and large-scale comparative assessment, followed by targeted validation or adjustment by human experts—particularly in high-stakes regulatory or policy-making contexts.

6. References

[1] Breiman, L. (2001) ‘Statistical modeling: The two cultures’, *Statistical Science*, 16(3), pp. 199–231.

[2] Calleo, Y. and Pilla, F. (2025) ‘Real-Time AI Delphi: A novel method for decision-making and foresight contexts’, Futures, 174, 103703. https://doi.org/10.1016/j.futures.2025.103703

[3] Clemen, R.T. and Winkler, R.L. (1999) ‘Combining probability distributions from experts in risk analysis’, *Risk Analysis*, 19(2), pp. 187–203.

[4] Cooke, R.M. (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: Oxford University Press.

[5] Dalkey, N. (1969) *The Delphi Method: An Experimental Study of Group Opinion*. Santa Monica, CA: RAND Corporation.

[6] Dalkey, N. and Helmer, O. (1963) ‘An experimental application of the Delphi method to the use of experts’, *Management Science*, 9(3), pp. 458–467.

[7] Keeney, R.L. and Raiffa, H. (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.

[8] Kendall, M.G. and Smith, B.B. (1939) 'The problem of m rankings', *The Annals of Mathematical Statistics*, 10(3), pp. 275–287.

[9] Klein, G., Moon, B. and Hoffman, R.R. (2010) 'Making sense of sensemaking 1: Alternative perspectives', *IEEE Intelligent Systems*, 21(4), pp. 70–73.

[10] Koo, T.K. and Li, M.Y. (2016) 'A guideline of selecting and reporting intraclass correlation coefficients for reliability research', *Journal of Chiropractic Medicine*, 15(2), pp. 155–163.

[11] McGraw, K.O. and Wong, S.P. (1996) 'Forming inferences about some intraclass correlation coefficients', *Psychological Methods*, 1(1), pp. 30–46.

[12] Nóbrega, L., Martinez, L. F., Marschhausen, L., Lima, Y., de Almeida, M. A., Lyra, A., Barbosa, C. E., & de Souza, J. M. (2025). AI-Delphi: Emulating Personas Toward Machine–Machine Collaboration. *AI*, *6*(11), 294. https://doi.org/10.3390/ai6110294

[13] O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester: Wiley.

[14] Rowe, G. and Wright, G. (1999) 'The Delphi technique as a forecasting tool: Issues and analysis', *International Journal of Forecasting*, 15(4), pp. 353–375.

[15] Shrout, P.E. and Fleiss, J.L. (1979) 'Intraclass correlations: Uses in assessing rater reliability', *Psychological Bulletin*, 86(2), pp. 420–428.