Dempster-Shafer Evidence Fusion

A method for fusion of evidence from separate sources. The method keeps disagreement and ignorance in the output.

> **Method name:** Dempster-Shafer evidence theory
>
> **Short name:** Evidence fusion
>
> **Main use in history:** Sensor fusion and expert systems
>
> **Project rank in the survey:** First. The method extends the judge lineage in a direct line.

Start Here: The Short Version

This method helps a computer combine answers from several judges. A judge can be an AI model, a rule, or a person. Each judge looks at the same case and gives an answer. The method does not force one final score too early. The method keeps three things clear. One part shows support for an answer. One part shows support against an answer. One part shows the part that is still unknown.

What it can do

- It combines several AI answers into one careful answer. The result keeps strong points from each judge.
- It shows when the judges agree and when the judges fight. The team sees the fight at once and does not miss the warning.
- It gives a clear next step for each case. The team can act, ask for more review, or stop because the proof is too weak.

Where people can use it

- A fraud team screens a claim with several tools. The method sends weak cases to a human. The method also sends disputed cases to a human.
- A safety team checks an AI answer before the AI sends the answer. The team also checks the answer before the AI takes an action. A weak result stops the action in time.
- A review team uses several tools to score the same file. The method joins the scores into one fair result. The team uses one result for the final call.

Problems it can help solve

- One AI model can sound sure and still be wrong. A single score hides that risk from the team.
- A simple average can hide a strong disagreement. Two judges can fight and the average can still look calm.
- Teams often have no clear rule for the next step. Teams need a rule for when to trust an answer. Teams also need a rule for when to stop.

> **The rest of this brief gives the history, the method, the 2026 gap, and a first project plan.**

1. IDENTIFICATION

Arthur Dempster published upper and lower probabilities in 1967.

Glenn Shafer published *A Mathematical Theory of Evidence* in 1976.

Shafer built a full theory of evidence from the work of Dempster.

Clinicians used a related practice in the MYCIN system in the 1970s.

MYCIN attached a certainty factor to each rule conclusion.

Engineers later used evidence theory for sensor fusion and fault diagnosis.

IBM Watson combined scores from many analysis pipelines in 2011.

The Watson system used the DeepQA architecture from IBM.

DeepQA split each question across separate analysis pipelines.

Each pipeline scored candidate answers with a different method.

Trained statistical models combined the pipeline scores into one ranking.

Watson did not use Dempster-Shafer theory.

Dempster-Shafer theory does not form the base of Watson.

Watson does not form the base of Dempster-Shafer theory.

The two lines of work share one problem only.

Both lines combine evidence from separate sources into one result.

The methods inside the two lines differ from start to end.

Watson stands in this brief as a parallel example of evidence combination at large scale.

2. PURPOSE

Dempster-Shafer theory combines evidence from separate sources and keeps disagreement visible.

Each source assigns belief to sets of answers and reserves a share for ignorance.

The fusion rule produces a belief interval and a conflict value for each decision.

A team uses the interval and the conflict value to act, to escalate, or to abstain.

Probability forces a single number for each answer.

A single number hides a dispute between sources.

A single number also hides a total lack of evidence.

Evidence fusion separates the two cases in the output.

3. OPERATION

Evidence theory assigns mass to single answers, to groups of answers, and to the full set.

Mass on the full set states ignorance in an explicit form.

Belief sums all mass that supports a claim.

Plausibility adds the mass that does not contradict the claim.

The gap between plausibility and belief shows the remaining ignorance.

The Dempster rule combines mass functions from independent sources.

The conflict value measures disagreement between sources.

Figure 1: Fusion flow, three judges to one decision

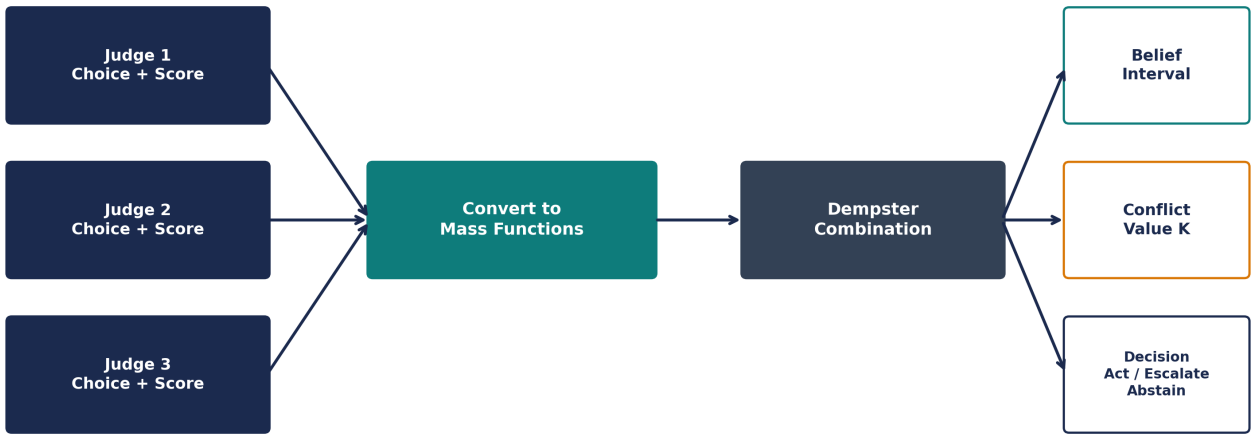

Figure 1: Three judges feed conversion and combination. Three outputs leave the combination block.

Procedure: fusion of two or more sources

1. Define the full set of hypotheses for the task.
2. Collect a mass function from each independent source.
3. Apply a reliability discount to each mass function.
4. Multiply the mass values for each pair of focal sets.
5. Assign each product to the intersection of the two sets.
6. Add all products on empty intersections. The sum gives conflict value K.
7. Divide the remaining mass by one minus K.
8. Read belief and plausibility for each hypothesis from the fused mass.
9. Route the case from belief, plausibility, and K.

OPERATION: WORKED EXAMPLE

Two judges assess the same fraud claim in this example.

Judge A and Judge B each report three mass values.

Source	Mass: Fraud	Mass: Not Fraud	Mass: Full Set (Ignorance)
Judge A	0.60	0.10	0.30
Judge B	0.50	0.20	0.30
Fused result	**0.76**	**0.13**	**0.11**

Products on empty intersections sum to 0.17. The sum gives conflict value K in this example.

Belief for Fraud equals 0.76 after fusion.

Plausibility for Fraud equals 0.87 after fusion.

The interval from 0.76 to 0.87 shows the remaining ignorance.

Figure 2: Belief interval and conflict gauge, worked example, hypothesis Fraud

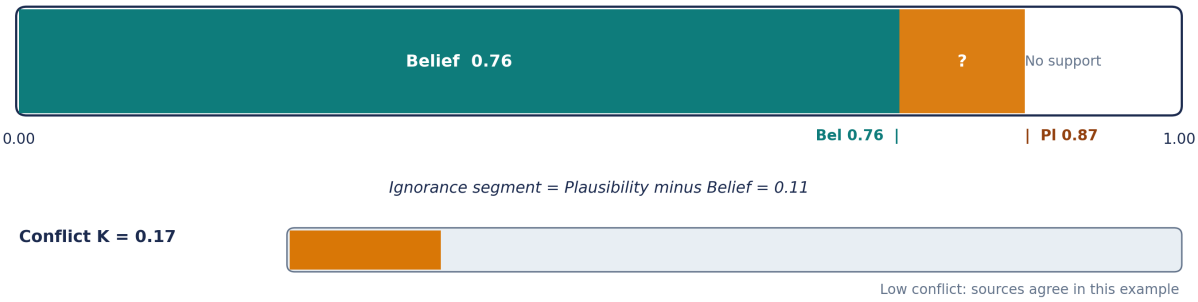

Figure 2: Belief segment, ignorance segment, and plausibility point for one hypothesis. The gauge shows conflict.

Decision routing from the output

A low conflict value and a high belief value support direct action.

A high conflict value signals a dispute between sources.

A large ignorance mass signals missing evidence.

The fusion layer routes each case to action, to review, or to data collection.

Core point: A standard probability average hides both signals. Evidence fusion reports both signals as numbers.

CONCRETE EXAMPLES: THREE TEAMS AT WORK

A fraud team screens one home health claim

A home health agency sends one claim for 4,800 dollars.

The claim lists twelve nurse visits in seven days.

A rules engine checks dates, codes, and visit counts.

The rules engine assigns 0.55 to Fraud, 0.05 to Not Fraud, and 0.40 to ignorance.

The notes describe nine visits and omit three billed visits.

The language model assigns 0.65 to Fraud, 0.10 to Not Fraud, and 0.25 to ignorance.

Fusion of the first two judges gives 0.83 for Fraud and 0.06 for Not Fraud.

Ignorance falls to 0.11 and conflict K stands at 0.09.

A history model then scores the mixed five-year record of the agency.

The history model assigns 0.35 to Fraud, 0.30 to Not Fraud, and 0.35 to ignorance.

The second fusion step raises K to 0.27.

The fused result gives 0.85 for Fraud, 0.10 for Not Fraud, and 0.05 for ignorance.

Belief for Fraud stands at 0.85 and plausibility stands at 0.90.

The team holds the payment and opens an investigation.

The fused record shows each judge vote and each conflict value to the investigator.

A safety team checks one AI action before release

A support agent drafts a refund action for one customer.

The draft sends 480 dollars to the customer.

The order total in the account stands at 240 dollars.

A policy checker reads the draft against the refund rules.

The policy checker assigns 0.70 to Safe, 0.10 to Unsafe, and 0.20 to ignorance.

A fact checker compares every figure in the draft with the account.

The fact checker finds the doubled refund figure.

The fact checker assigns 0.25 to Safe, 0.55 to Unsafe, and 0.20 to ignorance.

Fusion of the two checkers gives conflict K at 0.41.

The gate rule stops any action with K at 0.40 or higher.

The system blocks the refund and routes the draft to a human.

The human finds a duplicate line in the draft and corrects the refund to 240 dollars.

A review team scores one vendor file

A clinic group scores the security file of one software vendor.

A scanner checks 212 control answers in the file.

The scanner assigns 0.60 to Pass, 0.10 to Fail, and 0.30 to ignorance.

A document reader checks the policy documents behind the answers.

The reader assigns 0.50 to Pass, 0.05 to Fail, and 0.45 to ignorance.

Fusion of the first two tools gives K at 0.08.

One past audit shows an expired certificate.

The history check assigns 0.40 to Pass, 0.20 to Fail, and 0.40 to ignorance.

The fused result gives 0.84 for Pass, 0.09 for Fail, and 0.07 for ignorance.

Belief for Pass stands at 0.84 and plausibility stands at 0.91.

The team approves the vendor for a standard contract.

The team logs the expired certificate as a watch item for the next audit.

CONCRETE EXAMPLES: ONE ROBOT, THREE SENSORS

A delivery robot carries parts between two warehouse zones.

A camera, a lidar unit, and an ultrasonic sensor each assign mass values.

The frame of the task holds two answers: Path Clear and Path Blocked.

The behaviour bands of the robot

Band one covers belief for Path Clear at 0.90 or higher with K under 0.30.

The robot proceeds at full speed in band one.

Band two covers K at 0.30 or higher at any fusion step.

The robot slows, stops, and scans again in band two.

Band three covers belief for Path Blocked at 0.60 or higher.

The robot stops and calls a human operator in band three.

First scan: the sensors disagree

Sunlight through a dock door strikes the camera and the camera reads a dark shape across the lower frame.

The camera assigns 0.15 to Path Clear, 0.65 to Path Blocked, and 0.20 to ignorance.

The lidar assigns 0.75 to Path Clear, 0.05 to Path Blocked, and 0.20 to ignorance.

Fusion of camera and lidar gives conflict K at 0.50.

The pair result gives 0.58 for Path Clear and 0.34 for Path Blocked.

The ultrasonic sensor then reads the doorway space.

The ultrasonic sensor assigns 0.55 to Path Clear, 0.10 to Path Blocked, and 0.35 to ignorance.

The full fusion gives 0.75 for Path Clear, 0.21 for Path Blocked, and 0.04 for ignorance.

Belief for Path Clear stands at 0.75 after the full fusion.

The K value of 0.50 from the first pair still controls the behaviour band.

Band two applies. The robot slows, stops one metre before the doorway, and scans again.

Second scan: the sensors agree

The robot stands half a metre closer after the stop.

The glare angle changes and the camera separates shadow from frame.

The camera assigns 0.65 to Path Clear, 0.10 to Path Blocked, and 0.25 to ignorance.

The lidar assigns 0.70 to Path Clear, 0.05 to Path Blocked, and 0.25 to ignorance.

Fusion of the pair gives 0.88 for Path Clear and K at 0.10.

The ultrasonic sensor assigns 0.60 to Path Clear, 0.05 to Path Blocked, and 0.35 to ignorance.

The full fusion gives 0.95 for Path Clear, 0.02 for Path Blocked, and 0.03 for ignorance.

Band one applies. The robot proceeds through the doorway at full speed.

A pallet jack stands at the side of the doorway, clear of the path.

The robot logs both scans, all mass values, and both K values in the trip record.

Conflict changes the behaviour of the robot in direct steps.

4. HISTORY

Figure 3: Timeline, origin, main use, decline, and the 2026 gap

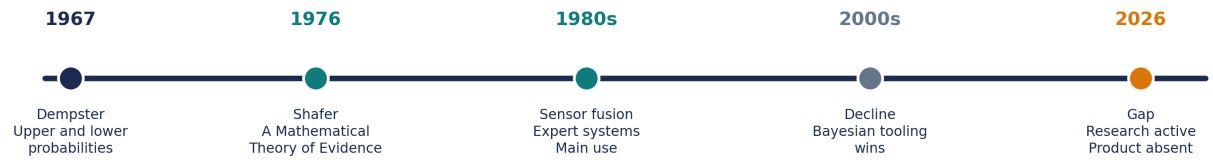

Figure 3: Origin, main use, decline, and the present gap in one line.

Teams used evidence theory in sensor fusion programs in the 1980s.

Teams also used evidence theory in expert systems in the 1980s.

Bayesian methods offered cleaner axioms and stronger software tools.

Bayesian methods won the tooling contest in universities and in industry.

Textbooks presented a famous counterexample for fusion under extreme conflict.

Zadeh described the counterexample. The example bears the name of Zadeh in the literature.

The example showed a strange result after fusion of two nearly opposed sources.

Tool makers built libraries for probability and left evidence theory in journals.

Software support for evidence theory stayed small for two decades.

Research teams averaged model outputs and discarded disagreement data for two decades.

5. THE 2026 GAP

Modern teams run many models and agents on the same task.

Product teams still average scores or take a majority vote.

Research teams published evidential deep learning papers in 2025.

One research line fused evidence masses inside medical image models.

Other research lines fused independent classifiers for multimodal tasks in 2026.

No mainstream product fuses outputs from language model judges with this method in 2026.

No mainstream product fuses outputs from agent judges with this method in 2026.

Standard ensemble averaging deletes the disagreement information.

The gap leaves a clear product opening for a small fusion service.

6. PROJECT PROPOSAL

Goal. The desired end state is one fusion service in daily use for typed judge panels. Any caller sends one case and receives one fused record. The record holds a belief interval, an ignorance value, a conflict value, and one decision word.

Proof target. The project proves one claim on labeled fraud dossiers. Dempster-Shafer fusion gives better calibrated decisions than the best single judge. The same fusion also beats plain score averaging on act decision errors at the same act rate.

Product at the end. The product is the DST Fusion Harness, version 1. The harness contains judge adapters, the fusion engine, the decision policy, and a full JSON audit trail. First users are judge panels in fraud screening and AI review work.

The project adds a fusion layer above existing judge calls.

Each judge sends a choice, a score, and a confidence value.

Judges include language models and Jev-style typed calls.

The service converts each output to a mass function.

The service fuses the mass functions and returns three values.

The three values are a belief interval, a conflict value, and a decision word.

The decision word takes one of three forms: act, escalate, or abstain.

A high conflict value routes the task to a human or to a larger model.

Components of the first build

- **Input adapter:** One adapter for each judge type. The adapter reads choice, score, and confidence.
- **Mass converter:** The converter builds a mass function from each judge output.
- **Discount store:** The store holds a reliability value for each judge.
- **Fusion core:** The core applies the Dempster rule and computes K.
- **Router:** The router maps belief, plausibility, and K to a decision word.
- **Audit log:** The log stores every input mass, every discount, and every fused result.

Steps for the first build

1. Build the fusion core in a single small library.
2. Check the core against hand calculations and known examples.
3. Connect three judge types to the input adapter.
4. Add discount learning from labelled cases.
5. Add a display for intervals, conflict, and source history.
6. Test the service on fraud dossiers with partial ground truth.
7. Measure agreement, escalation rate, and error rate against score averaging.

7. RISKS AND LIMITS

High conflict breaks the basic fusion rule in rare cases.

The design caps the effect of conflict and routes extreme cases to review.

Poor calibration in a judge produces poor mass values.

Regular calibration checks protect the fusion layer from silent drift.

Users need a simple display. Intervals confuse new readers at first sight.

Evidence fusion adds the two signals that score averaging deletes.

The math is small, old, and fully testable.

A focused prototype gives test results within weeks.

Sources: Dempster (1967) on upper and lower probabilities. Shafer (1976), A Mathematical Theory of Evidence. Survey report: Foundational Concepts series research, section 1, 2026-10-23.