

Quantization Loss in Safety-Critical Decision Systems

Greg Passmore

Abstract

Binary decision aids, that is, indicators that reduce a graded measurement to a category or a bit before an operator sees it, appear across safety-critical engineering. Radiation-therapy interlocks, reactor status lights, cockpit stall warnings, and combat identification systems all share this pattern. They also share a failure mode near the threshold, that is, the regime where the reduction discards the information a rational decision-maker needs to act. This paper treats the problem as one of information loss under quantization, examines four well-documented incidents as illustrative evidence, and argues for a design rule, that is, where a graded measurement is available, carry the confidence forward and present it alongside the decision rather than only the argmax. The four incidents help illustrate the argument; they do not attempt to prove it in a statistical sense.

Keywords: threshold quantization; information loss; graded posterior; operator interface; command and control; combat identification; safety-critical systems; human factors; fratricide.

1. Introduction

At 09:54 local on 3 July 1988, the tactical action officer aboard USS Vincennes watched a track symbol on the Aegis Large Screen Display. The symbol was tagged hostile. Two minutes earlier, the ship's SPY-1A radar had begun tracking a scheduled Iran Air A300 climbing out of Bandar Abbas, still in the published civil airway, still transmitting a civil Mode III IFF code, still climbing through roughly 12,000 feet. All of that graded evidence existed at the sensor. None of it was on the display. The crew fired two SM-2 missiles and killed 290 people, and the Navy investigation that followed named crew stress and scenario fulfillment as the cause [1].

The stress and the fulfillment are real. But underneath them sits an interface problem, that is, the ship's identification chain reduced a set of graded inputs to a discrete label before the operator saw them, and the reduction discarded evidence that could have caught the error. The same pattern appears in three other well-documented incidents, that is, the Therac-25 overdoses of the mid 1980s, the stall-warning inversion on Air France 447, and the stuck relief valve at Three Mile Island. In each case, graded evidence sufficient to distinguish the correct state existed somewhere in the sensing chain, and the interface presented to the operator did not carry it.

This paper treats that pattern formally. The claim is not that binary decision aids are always wrong, that is a straw target. Some decisions really are binary, and thresholds have to fire somewhere. The claim is narrower, that is, when a graded measurement is reduced to a bit before it reaches the operator, the reduction is asymmetrically lossy, discarding the most information near the threshold, where the decision is hardest. Section 2 makes that precise in information-theoretic terms. Section 3 recasts the same statement in Bayesian language. Section 4 walks through four incidents. Section 5 summarizes the safety-engineering literature that emerged from them. Section 6 states the design rule. Section 7 notes where the same failure reappears in modern autonomous and machine-learning systems.

2. The mathematics of decimation

Consider a sensor observing a physical state S , producing a noisy measurement $M = f(S) + N$, where N is a noise term with variance σ^2 . The measurement lands on an operator's display through some quantizer Q_K , which maps M into one of K discrete percept levels. The operator sees $P = Q_K(M; t)$, where t is the vector of thresholds that defines the quantization. The interface designer chooses K and t . The physics chooses σ .

Formally, the useful quantity is the mutual information between the true state and the operator's percept, that is, $I(S; P)$, in bits. By the data-processing inequality, $I(S; P) \leq I(S; M)$, with equality only when the quantization is lossless with respect to S , that is, when Q_K is a sufficient statistic for S given M . For any nontrivial noise process, no finite- K quantizer is sufficient, and the inequality is strict.

The quantity that matters for the operator is,

$$I(S; P) = \sum_s \sum_p \Pr(s, p) \log_2 \frac{\Pr(s, p)}{\Pr(s) \Pr(p)}.$$

For $K = 2$, that is, a binary reduction, this becomes an equation in one threshold t . For $K = 3$, two thresholds. For a continuous percept, the operator sees M directly and $I(S; P) = I(S; M)$, which is the information the physics allows. The gap between $I(S; P_{K=2})$ and $I(S; M)$ is the cost of the binary reduction, and it is not evenly distributed across the threshold's placement. Figure 2 shows the shape of that cost for a canonical case, that is, a binary state with unit separation between means and unit-variance Gaussian noise.

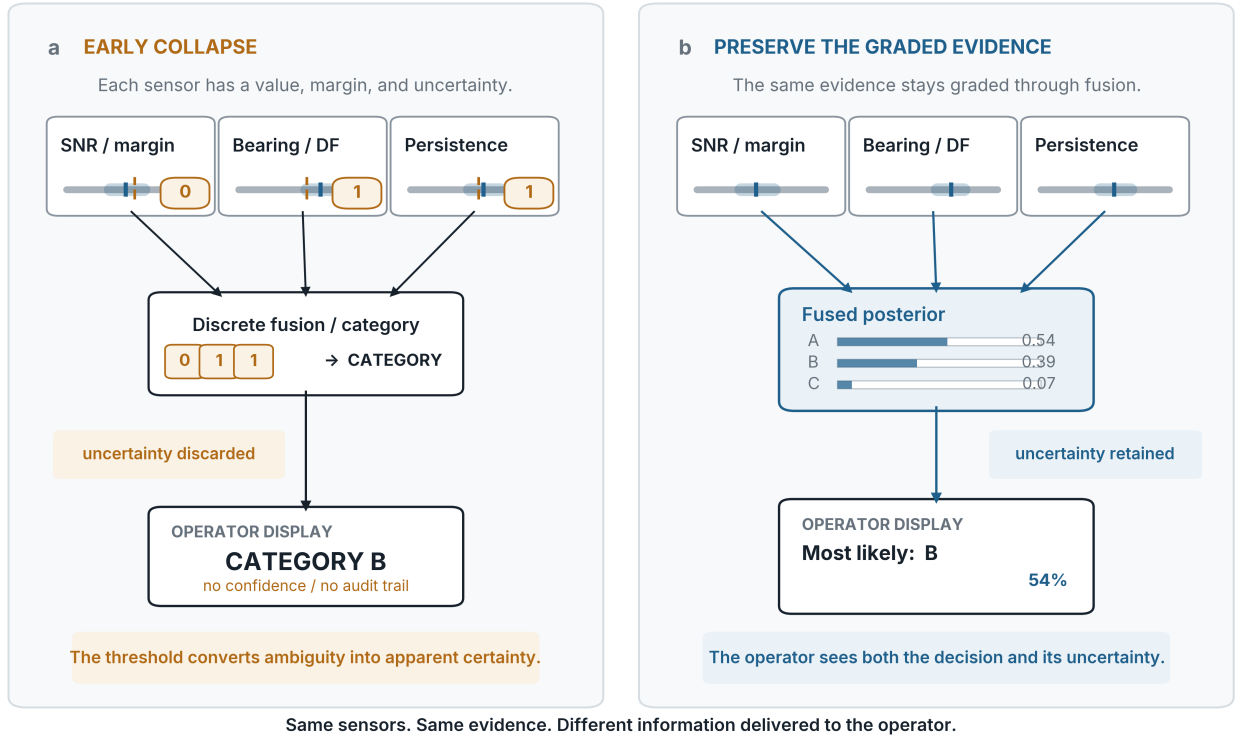


Fig. 1 - The same three sensors, the same measurements, two different interfaces. In (a) each sensor's graded reading, that is, value, margin, and uncertainty, is thresholded to a bit before fusion, and the operator sees only the argmax category with no confidence and no audit trail. In (b) the graded readings survive fusion, and the operator sees the posterior across candidate states along with the winning label. The two chains agree on the mean of the state estimate. They disagree on what happens at the margins, and the margins are where the decision matters.

Two observations from the figure carry the argument. First, the binary curve peaks at the midpoint between the two means at roughly 0.11 bits, against a continuous-percept limit of about 0.16 bits, that is, the binary reduction preserves roughly two-thirds of the available information under ideal threshold placement, and less as the threshold moves off the midpoint. Second, the three-level curve peaks near 0.14 bits and tracks the continuous limit far more closely than the binary curve, and it is less sensitive to threshold placement. These numbers are for the canonical unit-noise, unit-

separation case; the qualitative shape holds more generally, though the exact loss depends on the noise distribution, the separation between states, and the placement of the thresholds. The design cost of adding a middle category is small. The information cost of not adding it is not.

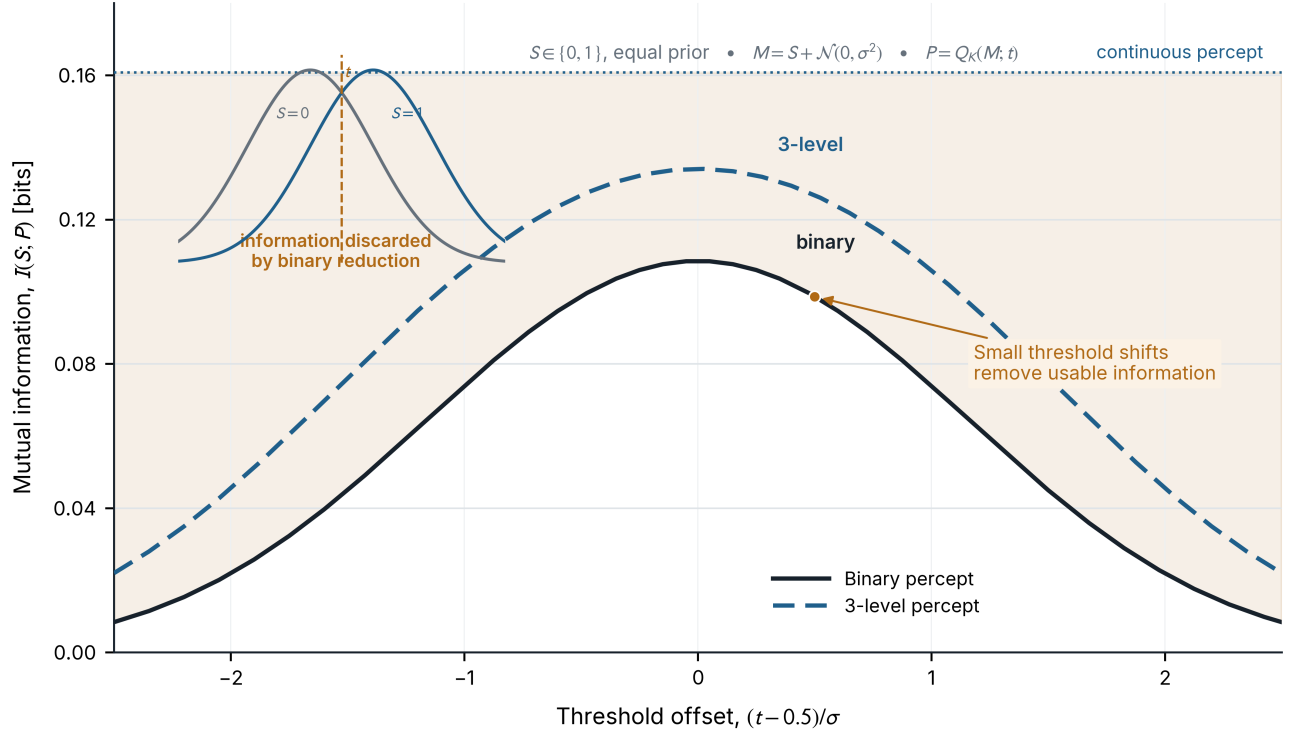


Fig. 2 - Mutual information between the true state and the operator's percept as a function of threshold offset, in units of the noise standard deviation. The shaded region between the curve and the continuous-percept limit is the information the interface throws away. The inset shows the two Gaussians for $S = 0$ and $S = 1$ under unit noise, with the threshold t crossing them near the intersection. A binary reduction peaks at the midpoint between the means and falls off quickly.

A three-level reduction tracks the continuous limit far more closely, and it tolerates threshold misplacement without collapsing.

This is the load-bearing technical point. Binary reduction is not just lossy, it is specifically lossy near the threshold, that is, at the operating regime where the sensor is telling the operator "I am not sure". A binary display converts such moments into an apparent claim of certainty. The operator then acts on the certainty, not on the underlying evidence. The four incidents in Section 4 illustrate what that looks like in practice.

3. The Bayesian view

The same statement rewrites in Bayesian language. Given prior $\Pr(S)$ and likelihood $\Pr(M | S)$, the posterior over the state after observing the measurement is,

$$\Pr(S | M) \propto \Pr(M | S) \Pr(S).$$

A rational operator acting on this posterior chooses the action that minimizes expected loss. In the interior of the state space, that is, where the posterior is concentrated on one hypothesis, the action is clear and a binary display is adequate, since the graded posterior and its argmax agree. At the margins, that is, where the posterior is diffuse across two hypotheses, the action depends on the shape of the loss function, and the argmax alone throws away the shape.

The Vincennes case illustrates the point. Before the missile release, the graded evidence available at the sensors, that is, the SPY-1A altitude track, the Mode III civil IFF returns, and the correlation with published civil airway schedules, was consistent with a civil airliner; the identification chain nevertheless produced a hostile label. Reconstructing the exact posterior across identity hypotheses at that moment is not possible from the public record, but the qualitative point is well-supported, that is, the evidence supported at least two identity hypotheses with materially different consequences, and the loss function for the two is not symmetric. A rational operator with the graded posterior in view chooses differently from an operator whose display has already argmaxed the posterior into a discrete label. The identification chain performed the argmax. The operator acted on it. The graded posterior existed and was not shown [2].

This is the reason a confidence indicator matters, even when the underlying decision is binary. The operator does not need to override the machine's classification, in fact the machine is usually right. What the operator needs is a channel for the graded moments, that is, the ones where the machine's classification is close to a coin flip, so the operator can slow the decision down, ask for corroboration, or hold. A binary display gives the operator no such channel. Every state looks like every other state, and the operator learns to trust it or to distrust it wholesale.

4. Four cases

The four cases below cover four different collapse points, that is, four different places in the sensing chain where graded evidence was reduced to a bit before it reached the operator. In each case the collapse is what the accident reports name as the load-bearing failure, though the human-factors literature is more direct about it than the incident investigations are.

4.1 Therac-25

The Therac-25 was a medical linear accelerator built by Atomic Energy of Canada Limited in the early 1980s. It delivered either a high-energy electron beam directly, or a lower-energy beam through a beam-flattening filter, depending on operator selection. A software interlock was supposed to guarantee that the filter was in place whenever the high-energy beam was selected. Between 1985 and 1987, at least six patients received massive radiation overdoses. Three died [3].

The interlock was a binary flag, that is, "safe to fire" or "not safe to fire", computed from two subsystems that maintained separate views of the machine's mode. A race condition let one subsystem update its view faster than the other, and for a brief window the interlock reported "safe" while the machine was configured to fire the high-energy beam without the filter. The graded information was present in principle, that is, the two subsystems held different views of the machine state, and neither view alone was authoritative. The interlock collapsed that internal disagreement to a single bit before it reached the operator. The operator saw the ready light and pressed the fire key.

The lost quantity was internal-state disagreement, that is, the fact that the two subsystems held different views of the machine at the moment of firing.

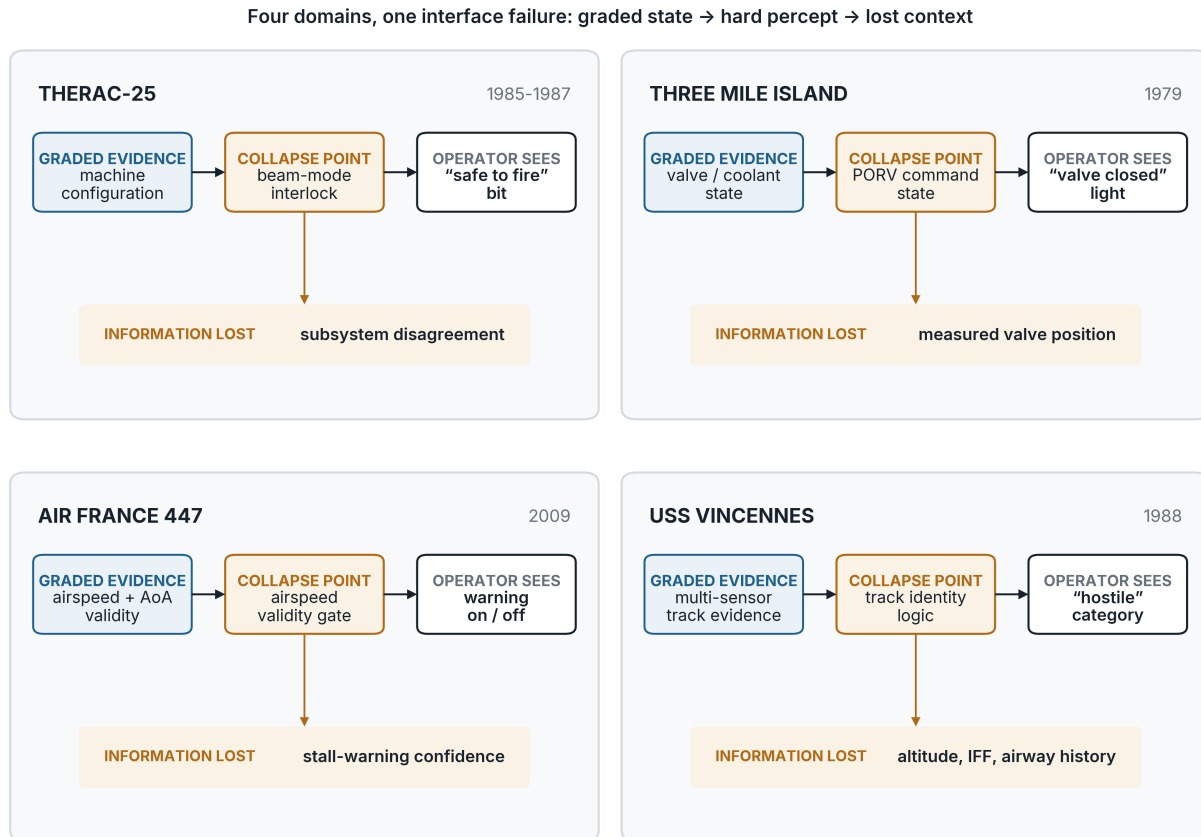
4.2 Three Mile Island

At 04:00 on 28 March 1979, a pilot-operated relief valve on the pressurizer of Unit 2 opened as designed to relieve overpressure. It should have closed within seconds. It stuck open, and it stayed open for two hours and twenty minutes while the reactor lost coolant [4].

The control-room indicator for the valve reported command state, not measured state, that is, the light showed whether the operators had commanded the valve closed, not whether it was actually closed. It was, in the strict sense, working correctly. It had been commanded closed. The valve, physically, was open. Direct position sensing on the valve stem was technically feasible, and its absence was among the deficiencies the Kemeny Commission and subsequent industry

practice specifically addressed [4]. Operators diagnosed the accident wrong for hours, and the reactor's fuel partially melted.

The lost quantity was measured valve position, that is, the actual physical state of the valve, which differed from its commanded state.



The recurring failure is not the sensor. It is the premature reduction of sensor state before operator inspection.

Fig. 3. Four incidents, four domains, one interface failure. In each case, graded evidence at the sensor was reduced to a hard percept before it reached the operator, and a specific quantity was lost at the collapse point, that is, subsystem disagreement in the Therac-25 interlock, measured valve position at Three Mile Island, stall-warning confidence on Air France 447, and the fused identity posterior aboard USS Vincennes. The recurring failure is not the sensor. It is the premature reduction of sensor state before operator inspection.

4.3 Air France 447

On 1 June 2009, Air France 447 entered a high-altitude stall over the South Atlantic and did not recover. All 228 people on board died. The immediate trigger was pitot-tube icing, which invalidated the airspeed reading. The aircraft's stall-warning logic gated the warning behind an airspeed-validity flag, that is, the warning sounded only when the computed airspeed was considered valid [5].

At the point of stall entry, the airspeed briefly fell below the validity threshold, and the stall warning cut off. When the pilots pushed the nose down and airspeed recovered enough to be valid again, the warning came back on. Every time they did the right thing, the alarm sounded. Every time they pulled up further into the stall, it went quiet. The graded information, that is, the confidence with which the airspeed reading could be trusted, existed inside the flight computer. It was reduced to a bit at the alarm gate, and the bit inverted the operator feedback.

The lost quantity was the airspeed-validity confidence, that is, the graded reliability of the airspeed reading that the stall warning depended on, presented to the crew as a hard on/off flag rather than as a degraded but persistent stall indication.

4.4 USS Vincennes

The Vincennes case sits at the top of this paper. The relevant technical detail, for this section, is that the ship's identification chain fused several graded inputs into a single track-identity label, that is, the SPY-1A altitude and climb-rate track, the Mode II and Mode III IFF returns, and the correlation with published civil airway schedules. The altitude track showed a climb, the Mode III return was consistent with a civil transponder, and the airway correlation was clean. A Mode II military return, associated with an Iranian F-14 known to be at Bandar Abbas, appeared in the identification chain and was correlated to the track by the fusion logic; the exact mechanism of that correlation has been debated in the subsequent literature [6].

The identification chain reduced the inputs to a single track label, that is, hostile, and the label went on the Large Screen Display. The tactical action officer and the captain acted on the label.

The lost quantity was the fused identity distribution, that is, the confidence-weighted picture across candidate identities, which was consistent with a civil airliner and inconsistent with an attack profile in several channels, but which the display did not carry.

5. What the literature settled on

Three lines of work in the safety-engineering literature bear on this pattern. Nancy Leveson's analysis of the Therac-25 is the canonical treatment in software safety, and its recommendations center on hazard analysis, safe design of software interlocks, and independent monitoring [3]. Charles Perrow's account of normal accidents frames tightly coupled interactive systems as prone to failures the operator cannot diagnose from the interface they are given [7]. James Reason's treatment of human error emphasizes that an operator's ability to reason about a graded situation depends on the interface preserving that graded structure [8]. None of these authors states the specific design rule this paper argues for; the rule below is my synthesis of the shared theme running through their work.

That rule is simple to state and hard to follow, that is, where a graded measurement is available and the operator has time to use it, do not present the operator with only the binary output. Present the graded evidence alongside. The alarm is the argmax. The confidence is the shape of the posterior. Both belong on the display when both are computable.

The rule is hard to follow because graded displays are more expensive to build, harder to train against, and, on a good day, tell the operator less. The trade only pays off on the bad days, and the bad days are rare. The four incidents above are consistent with the argument that the trade is worth making; they do not prove it in a statistical sense, and no single accident should.

6. The general principle

The principle underneath these cases is a design rule about interfaces, not about sensors or algorithms. It states, that is, the operator's display should preserve as much of the graded posterior as the operator can act on, and the collapse to a decision should happen as late in the chain as possible, ideally in the operator's own reasoning rather than in the interface that feeds it.

Specifically, this rule has three consequences. First, where a threshold operator in a safety chain is fed by a graded measurement, the interface should carry a confidence indicator alongside the binary output, that is, a graded companion signal that shows the distance from the threshold and the reliability of the measurement. Second, where a category label is produced by a fusion or classification algorithm, the interface should carry the graded evidence that fed it, so the

operator can audit the classification against its own inputs. Third, where an indicator reports command state rather than measured state, it should be labeled as such, and a measured-state indicator should be provided where technically feasible.

None of these is free. Each one adds display real estate, training overhead, and cognitive load. The incidents in Section 4 are consistent with the argument that the cost is worth paying in safety-critical settings, and the associated failure mode does not appear randomly across operating regimes. It concentrates near the threshold, which is where the operator most needs the display to reflect the graded state of the underlying evidence.

One residual is worth calling out, that is, some decisions really are binary and have to fire on a threshold. A reactor scram is binary. A missile release is binary. The design work in those cases is not to make the decision graded, it is to make sure the threshold is set with the graded picture in view, that the operator can see how close to the threshold the state currently sits, and that the confidence in the state estimate is on the display alongside the state itself. The threshold fires when it fires. What the operator needs is the graded context around it, up to and including the moment of firing.

7. The same failure in modern systems

The graded-to-binary collapse reappears in modern autonomous and machine-learning systems, in a form that is easy to miss because the sensors and the classifiers are more sophisticated than they were in 1988. A modern threat-classification network produces a soft output, that is, a probability distribution over class labels. The distribution is graded, and it carries the same information a fused sensor posterior would carry. The failure appears at the next stage, that is, the soft output is thresholded to produce a hard label, and the hard label is what the operator sees.

The current design pattern for this is calibrated confidence, that is, the classifier's output is calibrated so its reported probabilities match the empirical frequency of correctness at that probability. Calibration is necessary and it is not sufficient. A well-calibrated 0.6 posterior is still a graded quantity, and a display that shows the operator only the argmax has still discarded it. The argument of this paper applies to interfaces of this shape regardless of whether the underlying machine is a fusion filter, a Bayesian estimator, or a deep network.

The specific new failure mode in machine-learning systems is that the argmax hides the shape of the posterior across the whole class space, not just the closest competitor. A network that is 0.55 on hostile and 0.42 on civil is telling the operator something different from a network that is 0.55 on hostile and 0.44 on unknown. Both argmax to hostile. The graded output distinguishes them, and the graded output rarely reaches the operator's display.

8. Closing

Binary indicators are cheap, and they are usually right. The four incidents examined here are not evidence against binary indicators. They are evidence that when a binary indicator is wrong, it tends to be wrong in a specific way, that is, near the threshold, in the graded regime, where the operator most needs the display to carry what the sensors are actually saying. The design rule argued for here is to preserve the graded evidence and present it alongside the decision, so the operator can see the shape of the posterior, not just its argmax. The rule is not new, but these incidents are used to help illustrate what tends to happen when it is not followed.

References

- [1] W. M. Fogarty, Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988, U.S. Department of Defense, Aug. 1988.
- [2] U.S. General Accounting Office, Aegis-Class Cruisers: Combat System Deficiencies, GAO report on Aegis identification and combat-system performance, 1992.

[3] N. G. Leveson and C. S. Turner, An Investigation of the Therac-25 Accidents, IEEE Computer, vol. 26, no. 7, pp. 18-41, July 1993.

[4] J. G. Kemeny et al., Report of the President's Commission on the Accident at Three Mile Island, U.S. Government Printing Office, Oct. 1979.

[5] Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, Final Report on the Accident on 1 June 2009 to the Airbus A330-203 Registered F-GZCP Operated by Air France, Flight AF447 Rio de Janeiro-Paris, July 2012.

[6] G. I. Rochlin, Trapped in the Net: The Unanticipated Consequences of Computerization, Princeton University Press, 1997.

[7] C. Perrow, Normal Accidents: Living with High-Risk Technologies, Basic Books, 1984.

[8] J. Reason, Human Error, Cambridge University Press, 1990.