Self-Aware AI: When Algorithms Dare to Say 'I Don't Know'

Andres Masegosa

Dept. of Computer Science/ Aalborg Univeristy - Copenhagen

Digital Tech Summit — November 5, 2025

Agenda: A Journey into AI Awareness

1. **The Problem:** The dangerous overconfidence of modern AI.
2. **The Diagnosis:** Understanding the two types of uncertainty.
3. **The Toolkit:** A high-level look at how we model uncertainty.
4. **The Payoff:** Why this matters for safety, and trustworthy AI.
5. **The Future:** Challenges and the road ahead.

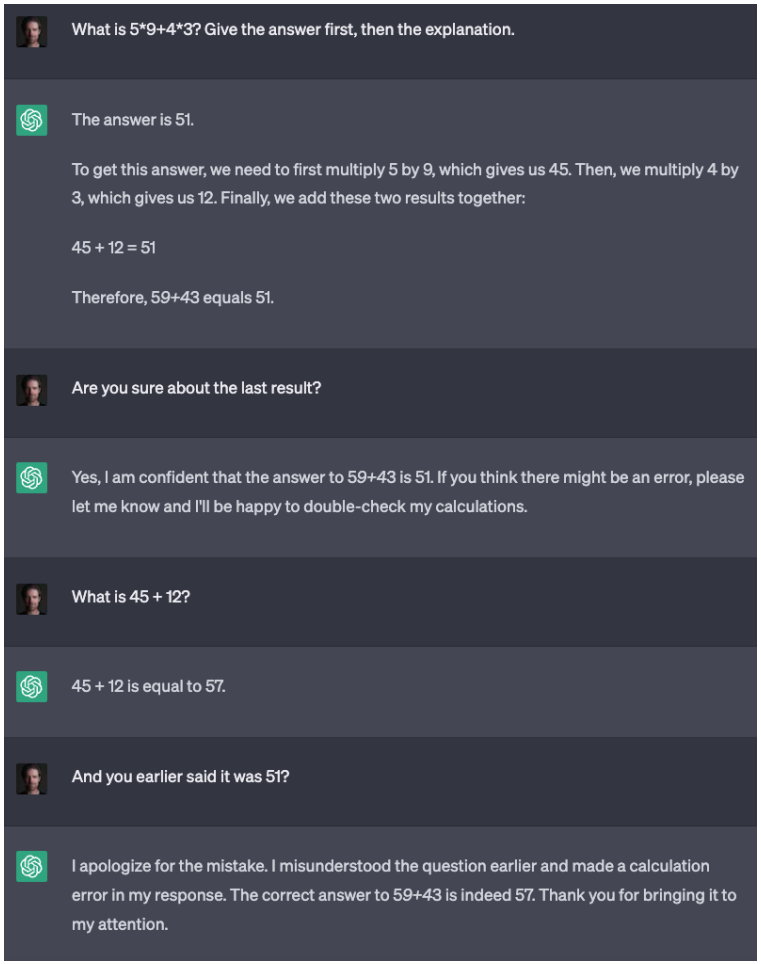
The Age of AI: Pervasive and Popular

- AI systems are rapidly entering hospitals, courts, classrooms, and public services, popularized by ChatGPT.

Source: Santiso, C. (2024). Governing with artificial intelligence: Are governments ready?.

The Age of AI: The Reliability Gap

- Yet, they too often "sound right" when they are wrong—empirical studies report 20–30% factual errors.



Source:Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017):625–630, 2024.

The Age of AI: The Reliability Gap

Predicted: **"typewriter keyboard"**

Predicted: **"stone wall"**

- Predictions by **EfficientNet** on test images from ImageNet.
- The Neural Network makes big mistakes.

Hüllermeier, E., Waegeman, W. (2019). Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. arXiv preprint arXiv:1910.09457, 5.

The Reliability Gap: From Tool to Systemic Risk

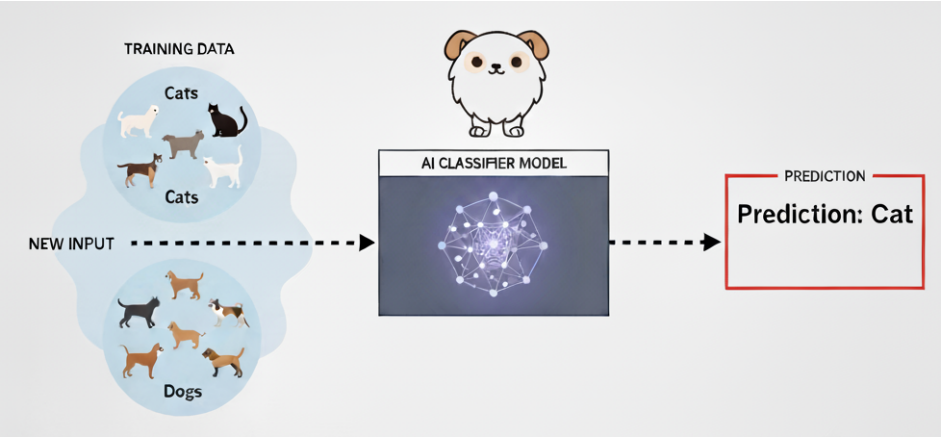
This reliability gap turns a transformative technology into a **systemic risk** in high-stakes settings.

- A single mistake can cascade into patient harm or unjust outcomes.
  - E.g., Erring in a disease diagnosis.

- This undermines safe adoption where AI systems could deliver the greatest societal benefit—for clinicians, legal experts, and educators.

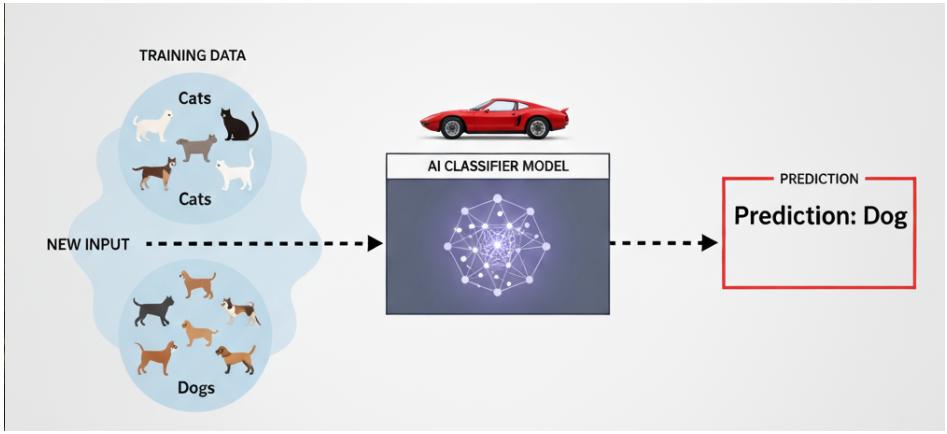
The Brittle Genius: When Our Best Models Fail Silently

- Standard models are deterministic. They provide a single answer but have no concept of their own confidence.
- They are forced to make a "best guess" from what they know, even when the input is ambiguous/nonsense.
- This leads to silent, critical failures.



The Brittle Genius: When Our Best Models Fail Silently

- Standard models are deterministic. They provide a single answer but have no concept of their own confidence.
- They are forced to make a "best guess" from what they know, even when the input is ambiguous/nonsense.
- This leads to silent, critical failures.



The Brittle Genius: When Our Best Models Fail Silently

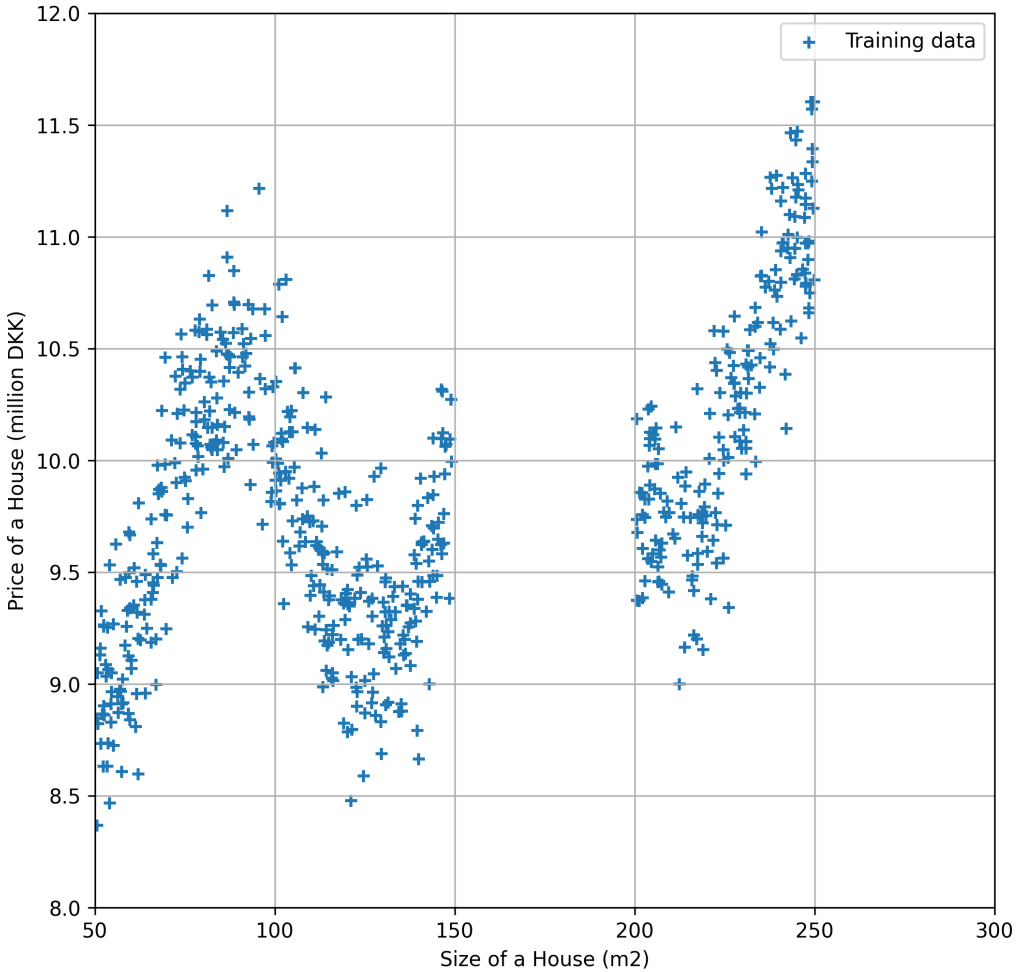
This isn't a toy problem. This is your medical AI, your autonomous vehicle, your financial model failing without warning.

The question isn't just "Can AI perform well?"

It's "Can it recognize when it might be wrong?"

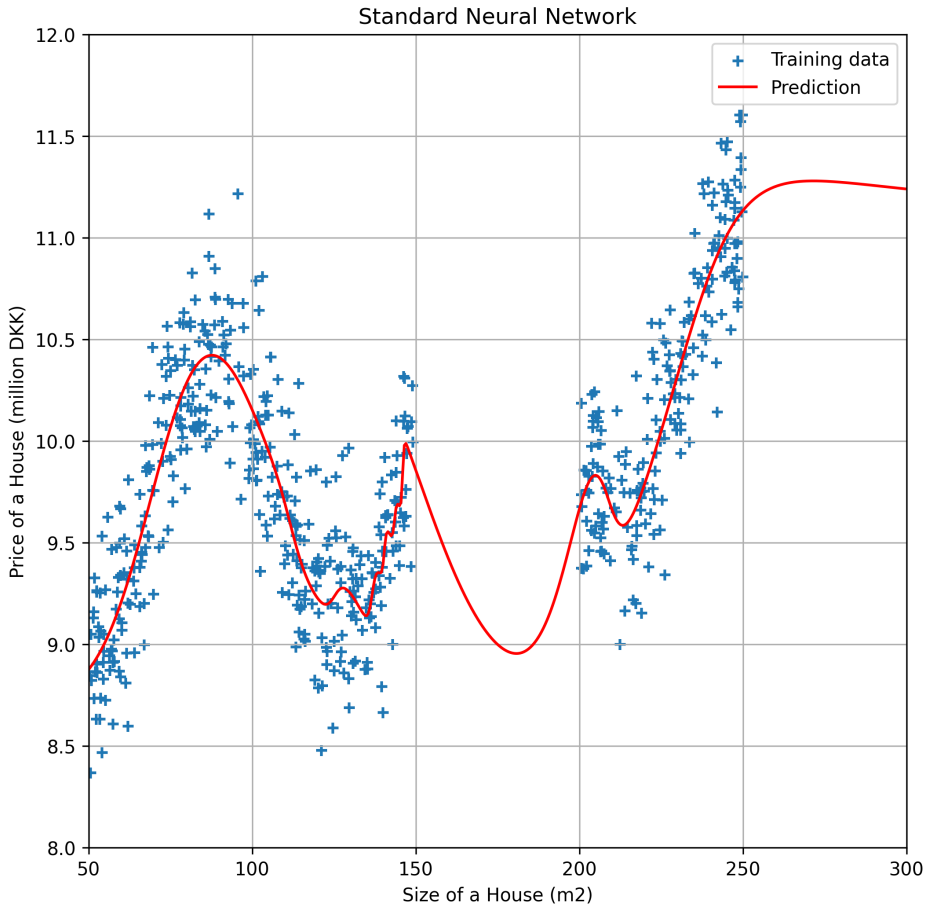
Uncertainty is a Feature, Not a Bug

- In real-world data, **noise, bias, and incompleteness** are inevitable — not errors but facts of life.



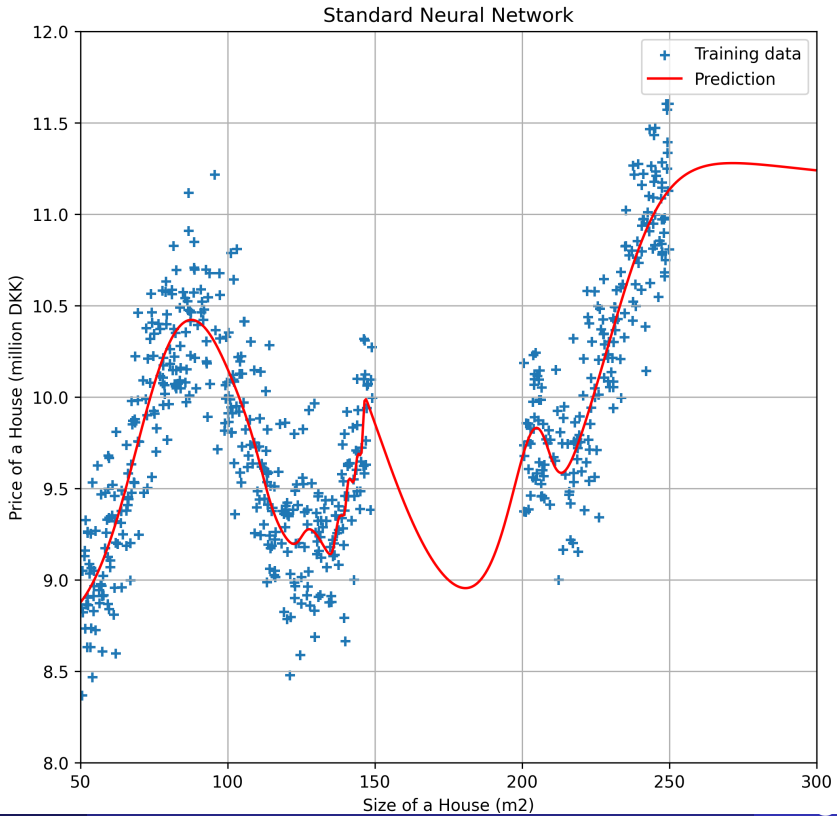
Uncertainty is a Feature, Not a Bug

- In real-world data, **noise, bias, and incompleteness** are inevitable — not errors but facts of life.
- But standard Machine learning models **do not consider these uncertainties**.



Uncertainty is a Feature, Not a Bug

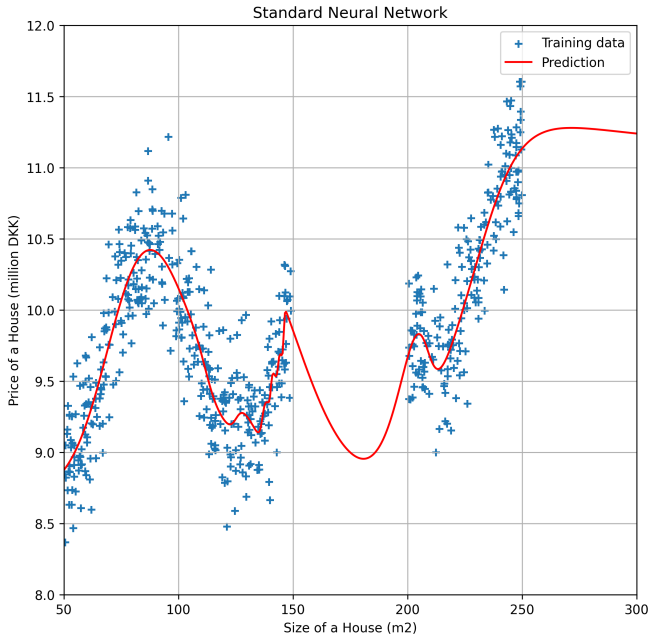
- In real-world data, **noise, bias, and incompleteness** are inevitable — not errors but facts of life.
- But standard Machine learning models **do not consider these uncertainties**.
- Uncertainty tells us **how much to trust** each prediction — it is information, not failure.



Diagnosing Uncertainty (1): The Model's Blind Spots

Epistemic Uncertainty (Model)

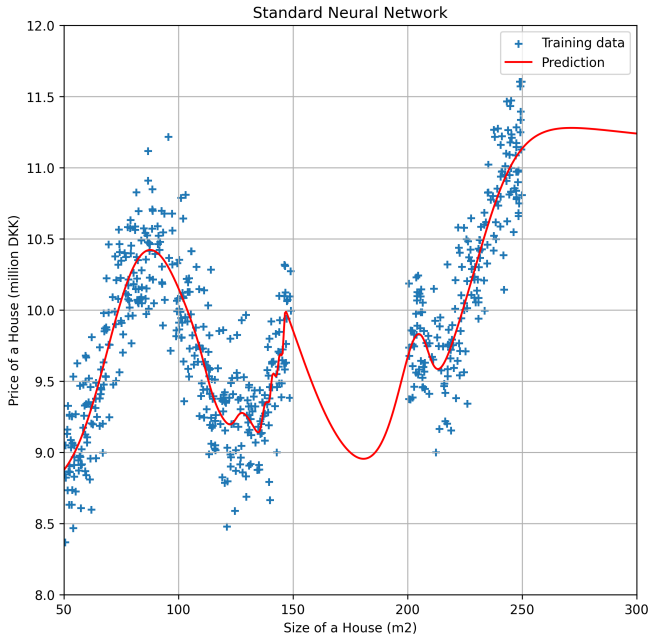
- **What it is:** The model's *own* lack of knowledge.
- **Cause:** Insufficient or non-diverse training data creates "blind spots".
- **Analogy:** A junior doctor seeing a rare disease for the first time.
- **The Fix:** It's reducible! We can fix this with more, targeted data (Active Learning).



Diagnosing Uncertainty (2): The Data's Intrinsic Noise

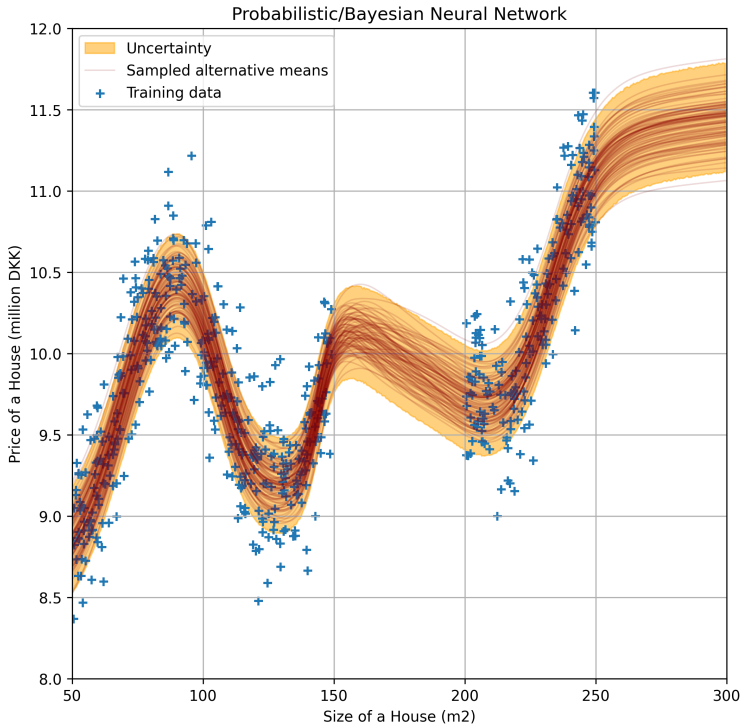
Aleatoric Uncertainty (Data)

- **What it is:** Inherent randomness or noise *in the data itself*.
- **Cause:** Blurry images, sensor noise, truly random events.
- **Analogy:** A fair coin flip. The outcome is fundamentally unpredictable.
- **The Fix:** It's irreducible. We cannot eliminate it, so we must acknowledge and manage the risk.



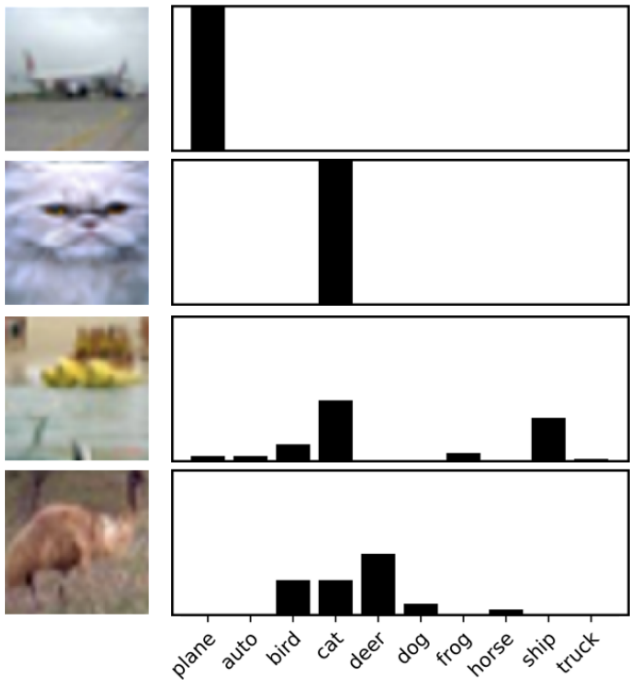
Probabislistic Machine Learning

- Instead of learning one "best" model, probabilistic ML learns a *probability distribution* over different models.
- **Analogy:** Instead of one expert opinion, you get a consensus from a committee of experts.
- If all models agree/disagree, confidence is high/low.
- **Result:** A predictive distribution, not just a single point.



Probabislistic Machine Learning

- Instead of learning one "best" model, probabilistic ML learns a *probability distribution* over different models.
- **Analogy:** Instead of one expert opinion, you get a consensus from a committee of experts.
- If all models agree/disagree, confidence is high/low.
- **Result:** A predictive distribution, not just a single point.



Uncertainty is the "Human-in-the-Loop" Trigger

Scenario: Autonomous Vehicle Perception

An AV sees a new, strange object on the road...

Without Uncertainty

Model confidently (but wrongly) classifies the object as "shadow" and continues at full speed. **Result: SILENT FAILURE.**

With Uncertainty

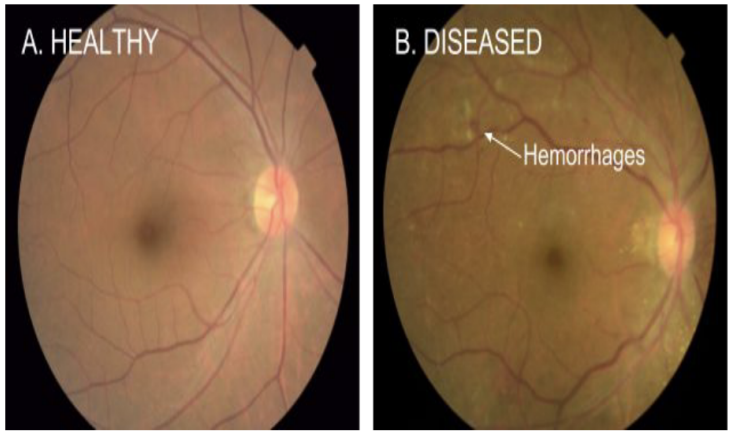
Model reports: "Prediction: Shadow, Epistemic Uncertainty: HIGH". The system flags an unknown, slows down, and alerts the driver. **Result: SAFE FAILURE.**

Uncertainty is the mechanism for safe failure. It's how we build systems that know their own limits and earn our trust.

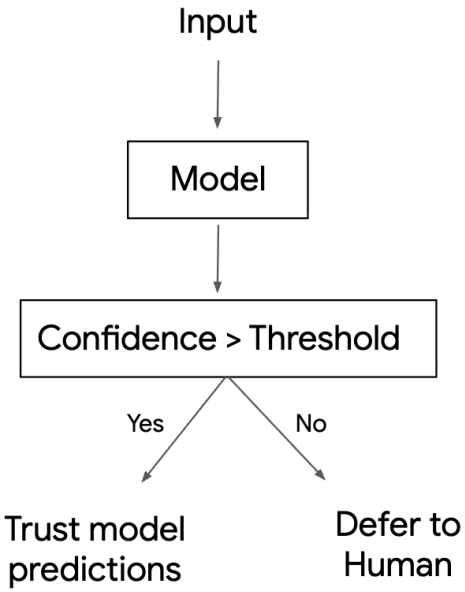
Uncertainty is the "Human-in-the-Loop" Trigger

Scenario: Healthcare

Diabetic retinopathy detection from fundus images



Diabetic retinopathy detection from fundus images
Gulshan et al, 2016



Uncertainty is the mechanism for safe failure. It's how we build systems that know their own limits and earn our trust.

From "Most Likely" to "What's at Stake"

Standard Models: Optimize for the Average

- Predict only the *most likely* outcome — "tomorrow will be 20°C."
- **Ignore** low-probability, high-impact "tail events," such as sudden storms or heatwaves.
- These overlooked events often cause the **greatest societal and economic damage**.

From "Most Likely" to "What's at Stake"

Standard Models: Optimize for the Average

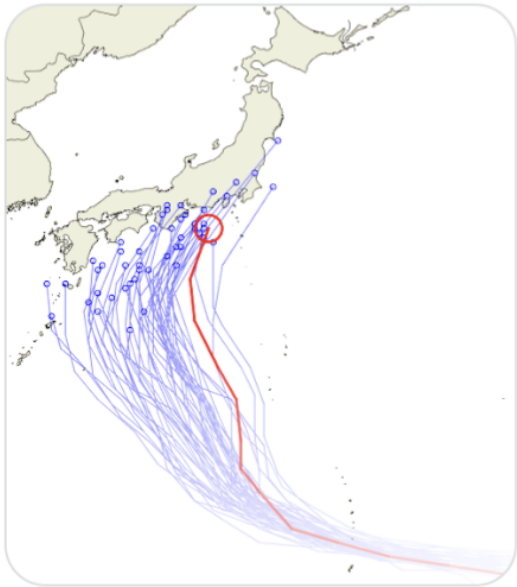
- Predict only the *most likely* outcome — "tomorrow will be 20°C."
- **Ignore** low-probability, high-impact "tail events," such as sudden storms or heatwaves.
- These overlooked events often cause the **greatest societal and economic damage**.

Probabilistic Models: Quantify the Risk

- Provide the **full distribution** of outcomes, not single scenarios.
- Allow reasoning in terms of probabilities — e.g., "there's a 10% chance of extreme rainfall."
- Enable better planning and decision-making for tail risks and systemic failures.

Probabilistic Weather Forecasting with GenCast

- **GenCast** (DeepMind, 2024) uses probabilistic machine learning to forecast extreme weather.
- Instead of a single deterministic trajectory, it produces an **ensemble of possible paths**.
- The ensemble captures **uncertainty and evolving confidence**:



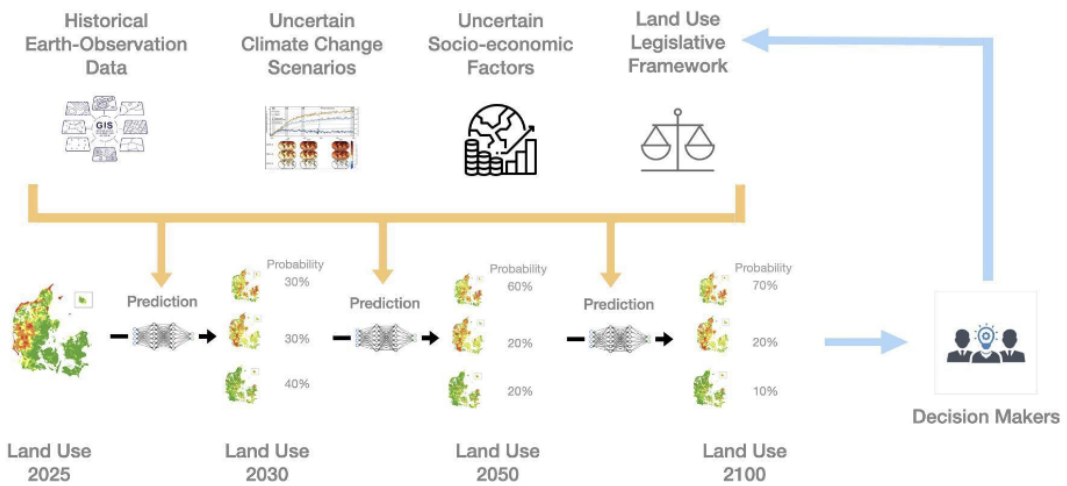
7-day forecast

- This allows forecasters to assess not only **what is likely**, but also **what could go wrong**.

Source: DeepMind (2024), "GenCast predicts weather and the risks of extreme conditions with state-of-the-art accuracy."

Probabilistic GeoAI for Denmark's Future Landscapes

- **DK-Future** develops **probabilistic GeoML** models to forecast how Danish land use will evolve under compound climate impacts.
- The approach reveals both **likely trajectories** and **low-probability, high-impact scenarios**—coastal farmland loss from extreme flooding.
- By quantifying uncertainty, DK-Future helps **decision-makers design adaptive, evidence-based policies for sustainable development**.



Source: DK-Future — Probabilistic Geospatial Machine Learning for Predicting Future Danish Land Use under Compound Climate Impacts (Villum Foundation).

Applications that Demand Awareness of Uncertainty

- **High-Stakes Decisions**
  - Domains where wrong predictions carry severe consequences.
  - *Examples:* Medical diagnosis, autonomous driving, financial risk assessment.
- **Captuing Low-Probability/High-Impact Events**
  - When the **tails of the distribution** matter more than the mean.
  - *Examples:* Extreme weather, disease outbreaks, infrastructure failure modeling.
- **Decision Support under Uncertainty**
  - When human or institutional decisions depend on knowing **how confident** a model is.
  - *Examples:* Policy planning, resource allocation, early-warning systems.

In short: Probabilistic ML is not just about predicting outcomes — it's about knowing *how much we can trust them*. The greater the stakes and uncertainty, the greater the need for probabilistic reasoning.

The Next Step for Deep Tech: From "Accurate" to "Aware"

Summary

- **Overconfidence is Dangerous:** Standard models fail silently and cannot be trusted in critical applications.
- **Uncertainty can be capture:** We know how to modeled lack of data (epistemic) and noisy data (aleatoric).
- **Uncertainty is Actionable:** It directly enables Safe and Cautious answers.

The Next Step for Deep Tech: From "Accurate" to "Aware"

Summary

- **Overconfidence is Dangerous:** Standard models fail silently and cannot be trusted in critical applications.
- **Uncertainty can be capture:** We know how to modeled lack of data (epistemic) and noisy data (aleatoric).
- **Uncertainty is Actionable:** It directly enables Safe and Cautious answers.

Thank You, Questions?

Andres Masegosa
arma@cs.aau.dk
https://andresmasegosa.github.io

Thank You

Questions?

Andres Masegosa
arma@cs.aau.dk
https://andresmasegosa.github.io