95% Accuracy Is Not All You Need

Kevin A. Bryan
U of T, Rotman

Joshua Gans
U of T, Rotman

Discussant: Yuk-fai Fong

Introduction

- This paper asks an important question: *When is a 95%-accurate AI truly useful for decision-making, and when might that 5% error be too dangerous?*
- When and how to optimally use AI?
- How to formalize a theoretical model for AI utilization that illustrates the key tradeoffs?
- Generalization of the model for other applications

Key Model Elements

- A decision-maker (DM) faces N possible states of the world, indexed by $\theta \in \Theta = \{1, 2, \ldots, N\}$. Each state is initially equally likely with probability $1/N$.
- The DM must choose an action $a \in \Theta$ (i.e. select one of the N states).
- The payoff for the chosen action a given the true state θ is defined as

$$u(a, \theta) = \begin{cases} H, & \text{if } a = \theta, \\ L, & \text{if } a \neq \theta \ , \end{cases}$$

  where $H > 0$ and $L < 0$.
- Thus H is the reward when action matches the state, and L (a negative value) is the penalty for choosing the wrong state.
- The DM also has the option to take no action and receive the outside option of zero

Key Model Elements: Tools for Guiding Decision Making

- **Verification (Human Effort)**: The DM can verify *with certainty* whether a state θ is the true state at cost c per test, one state at a time.
- **Statistical Method (Traditional test)**: At cost c_s, the DM can run a procedure S. This procedure returns the true state (accurately) with probability p, and with probability $1-p$ it returns *no information* (i.e., *it knows when it doesn't know*)
- **Black-box AI tool**: At cost c_α, the DM can use the AI tool. The AI always returns a single guess θ_α for which state is true. This guess is correct with probability q (where $q > p$), and wrong with probability $1-q$. *AI doesn't know when it's wrong*: overconfidence, hallucination

Main Result 1: When to Not Use AI

Proposition 1 (AI May Not Be Used Despite High Accuracy). *A necessary and sufficient condition for* not *using the AI in equilibrium is*

$$\max\{\pi_{a,g},\ \pi_{a,v},\ \pi_{s,a,g},\ \pi_{s,a,v}\} \ \leq\ \pi^*_{\text{non-AI}}.$$

Equivalently, letting

$$\begin{aligned}
T_{A,G} &= \frac{\pi^*_{\text{non-AI}} + c_\alpha - L}{H - L},\\
T_{A,V} &= 1 - \frac{H - c - c_\alpha - \pi^*_{\text{non-AI}}}{\beta(N-1)\,c},\\
T_{S,A,G} &= \frac{1}{H-L}\left(\tfrac{\pi^*_{\text{non-AI}} - \pi_{s,o}}{1-p} + c_\alpha - L\right),\\
T_{S,A,V} &= 1 - \frac{H - c - c_\alpha - \frac{\pi^*_{\text{non-AI}} - \pi_{s,o}}{1-p}}{\beta(N-1)\,c},
\end{aligned}$$

the AI will not be used if

$$q \ \leq\ \min\{\,T_{A,G},\ T_{A,V},\ T_{S,A,G},\ T_{S,A,V}\,\}.$$

In particular, when verification cost c is very large *(driving $T_{A,V}, T_{S,A,V} > 1$) and the cost of errors L is high, the DM may not use AI at all even if AI accuracy q is close to 1 and AI cost c_α is close to zero.*

- **Proposition 1 in words**: **When penalty for mistakes** $|L|$ is too large, **one must rely more heavily on verification or simpler statistical methods** to avoid catastrophic errors that even a very accurate AI sometimes makes.
- However, **if verification itself is also prohibitively expensive** (c large), then the DM cannot cheaply fix AI's black-box errors, making it **too risky to use AI at all**. This effect is magnified if the cost c_α of obtaining the AI guess is also high relative to the payoff improvement it can offer over the best non-AI strategy, but holds even if AI is free.

Main Results

Proposition 2 (AI With Verification or Statistical Pre-Screen). *Suppose* $q > \min\{T_{A,G}, T_{A,V}, T_{S,A,G}, T_{S,A,V}\}$ *so that an AI-based strategy is used in equilibrium. Then* AI alone *(meaning the strategy* (A, G) *is suboptimal if at least one of the following holds:*

1. *Verification is cheap enough compared to the potential cost of AI errors:*

$$c \leq \frac{(1-q)(H-L)}{1+(1-q)\beta(N-1)}.$$

2. *Statistical pre-screening* (S) *is cheap enough relative to* p, c_α, *and the cost of AI's errors. Formally,*

$$c_s \leq \max\Big\{p(1-q)(H-L)+pc_\alpha, \quad c_s < (1-q)(H-L)-c(1-p)[1+(1-q)\beta(N-1)]+pc_\alpha)\Big\}.$$

In either case, the DM uses AI in combination *with verification or with* S*, but never the AI by itself.*

Main Results

- **Proposition 2 in words**: Focus on the case that it is optimal to use AI. The DM **prefers to verify** AI's output or **use a statistical pre-screen** when the **costs of these safeguards are low** compared to the potential cost of AI errors.
- This explains why AI systems in **high-stakes domains** (healthcare, autonomous driving, cyber security) are **rarely deployed without human oversight** or complementary systems, even as their accuracy improves.

Main Results: Extensions

Highly generalizable Framework

- Adversarial Environment
- O-Ring Tasks (Agentic AI)
- Autonomous AI (No Human Oversight)
- Self-Correcting or Chain-of-Thought AI
- Multiple Agents (Committee) Voting on AI Advice
- Limited Liability
- Competition

Main Contributions

- Formalization: Allows systematic analyses of optimal and responsible AI adoption and utilization, capturing key tradeoffs
- Apart from high accuracy of AI, how costly errors are and whether the AI's errors can be detected or mitigated are equality important
- Value of knowing when you don't know and refraining from decision
- Highly generalizable: adversarial input manipulation, o-ring tasks, etc.

Some Potential Extensions

- Simplifying assumptions: human verification is perfect, statistic models know when they're uninformative, but AI models don't
  - Sometimes AI is used for risk management than a source of risk
  - How this model apply to content-generating LLMs which may not be substitutes for statistical models
- Exploration of regulatory implications?
- Extension to game-theoretic setting?
- How do we extend the model to analyze over-reliance on AI?

Main Contributions

- In sum, in the context of wide AI adoption,

  Bryan and Gans provide a timely framework
  showing that not just AI accuracy,
  but knowing AI's limits and the cost of its mistakes
  and augmenting AI usage with human judgment and statistic models
  to identify these mistakes and imitigating their impacts
  are equally important.

- Also, this paper will inspire many more theoretical work on AI.