Overcoming Algorithm Aversion: *Ex-Post* Human-in-the-Loop *Appeal* Process

Qiong Xia (presenting) [1] Geoff Tomaino [2] Theodoros Evgeniou [1] Klaus Wertenbroch [1]

[1]INSEAD [2]University of Florida

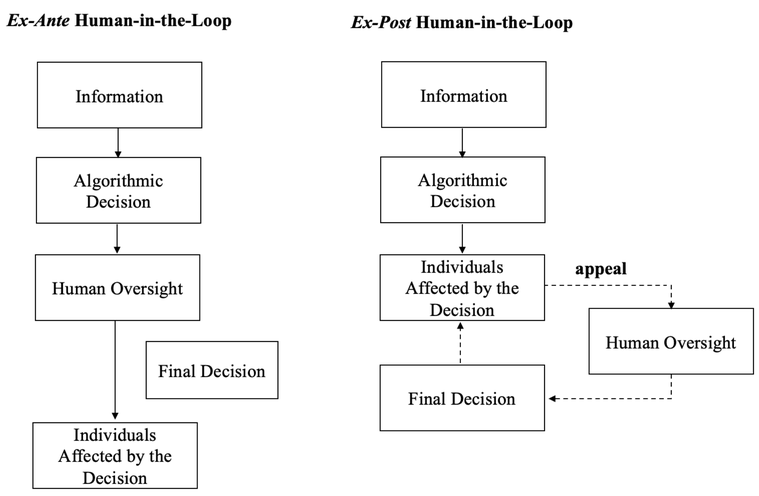
Broader context

- AI algorithms highly effective at prediction tasks
- Yet AI adoption remains slow:
  - $< 3\%$ of hospitals (Goldfarb, Taska, and Teodoridis 2020)
  - $< 2.5\%$ of worker roles (Babina et al., forthcoming)
  - McKinsey: global AI adoption rates plateaued since 2019
- Algorithm Aversion (e.g. Dietvorst, Simmons, and Massey 2015)
- Bring the human-in-the-loop of the algorithmic decision:
  - *Ex-ante* human oversight before reaching individuals affected by the decisions (Dietvorst, Simmons, and Massey 2018; Burton, Stein, and Jensen 2020; Sele and Chugunova 2022)
- Compromise the benefits of the algorithmic decision:
  - Lower accuracy (Sele and Chugunova 2022)
  - Introduce human bias (Tversky and Kahneman 1974)
  - Less reliable (Meehl 1954; Dawes, Faust, and Meehl 1989)
- How can we optimize the placement of human oversight within the algorithmic decision-making process?

Executive summary

- We propose an innovative human-in-the-loop approach: an *ex-post* human-in-the-loop upon *appeal* to maintain the benefits of algorithms and meet the regulatory concerns
- We focus on the preferences of individuals who are affected by the decisions between:
  - *Ex-ante* human-in-the-loop
  - *Ex-post* human-in-the-loop
- The aversion to algorithms makes individuals prefer the *ex-ante* over *ex-post* human-in-the-loop (Study 1)
- We nudge individuals to prefer *ex-post* over ex-ante human-in-the-loop by triggering their analytical thinking (Study 2)

Ex-post human-in-the-loop approach



Advantages of *ex-post* human-in-the-loop

- For firms:
  - Reduce costs
  - Scales of algorithm deployment
  - Enhance algorithmic machine learning
- For individuals affected by the decisions:
  - Offer a second chance for favorable outcomes

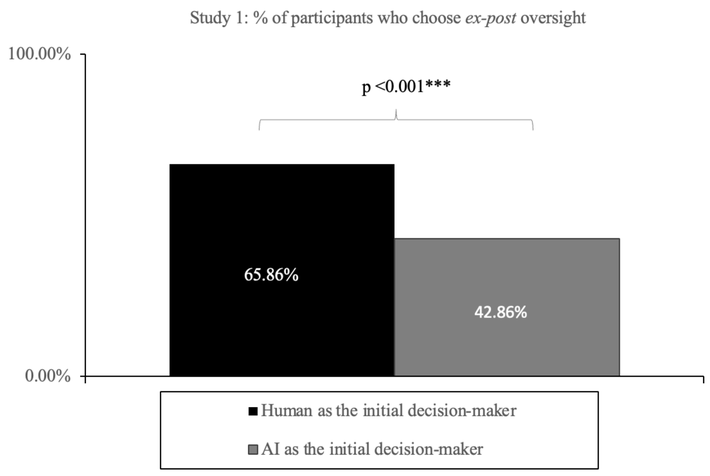
Studies overview

- Both studies are pre-registered
- Study 1:
  - N = 295 from Prolific, after excluding those who failed the attention check
- Study 2:
  - N = 974 from Prolific, after excluding those who failed the attention check

Study 1

- To address:
  - Are individuals aware of the advantages of *ex-post* oversight?
  - Preferences between *ex-ante* and *ex-post* human-in-the-loop algorithmic decisions
- Design:
  - A hypothetical bank loan application scenario
  - The banks use two options to decide to approve or reject the loan
  - 2-cell between-subject design, varying in the initial decision-maker: a human or an algorithm
- DV:
  Binary choice between Option 1 (*ex-ante* oversight) and Option 2 (*ex-post* oversight)

Results of study 1

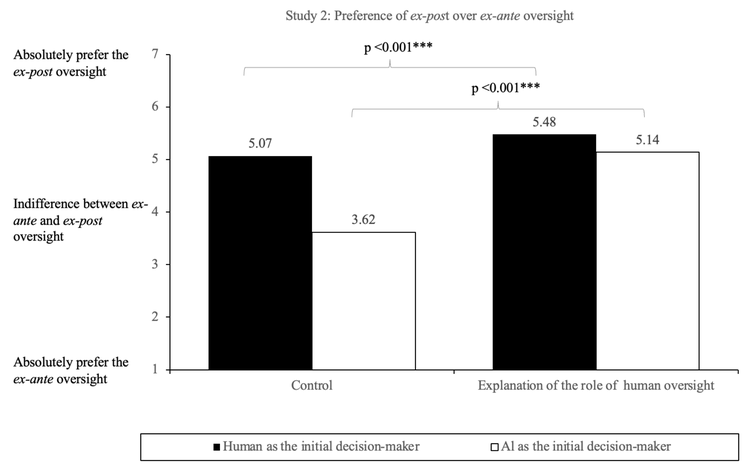


Study 2

- To address:
  - Can we nudge individuals to prefer *ex-post* over *ex-ante* human-in-the-loop?
- Design:
  - A hypothetical bank loan application scenario
  - The banks use two options to decide to approve or reject the loan
  - 2×2 between-subject design, varying in (1) the initial decision-maker: a human or an algorithm; (2) whether to explain the role of human oversight
  - Description of the human oversight: although such a revision could swing in your favor, it might also lead to an unfavorable outcome, especially if the initial decision was already favorable to you
- DV:
  - 7-point Likert scale (1 – Absolutely prefer *ex-ante* oversight, 4 – Indifferent between *ex-ante* and *ex-post* oversight, 7 – Absolutely prefer *ex-post* oversight)

Results of study 2

- ANOVA interaction: $F(1, 970) = 18.69$; $p < 0.001$, $\eta^2 = 0.02$



Next

- Driving preference: aim towards adopting the *ex-post* human-in-the-loop approach
- Potential mechanism why individuals do not exhibit a preference for *ex-post* human-in-the-loop

References

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. Forthcoming. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics*.

Burton, Jason W, Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A systematic review of algorithm aversion in augmented decision making." *Journal of Behavioral Decision Making* 33 (2): 220–239.

Dawes, Robyn M, David Faust, and Paul E Meehl. 1989. "Clinical versus actuarial judgment." *Science* 243 (4899): 1668–1674.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144 (1): 114.

———. 2018. "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them." *Management science* 64 (3): 1155–1170.

Goldfarb, Avi, Bledi Taska, and Florenta Teodoridis. 2020. "Artificial intelligence in health care? Evidence from online job postings." *AEA Papers and Proceedings* 110:400–404.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.

Sele, Daniela, and Marina Chugunova. 2022. "Putting a Human in the Loop: Increasing Uptake, but Decreasing Accuracy of Automated Decision-Making." *Max Planck Institute for Innovation & Competition Research Paper*, nos. 22-20.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *science* 185 (4157): 1124–1131.

qiong.xia@insead.edu