Behavioral Economics

Lecture 3: Updating with AI Advice

Daniel Martin
danielmartin@ucsb.edu

Thanks to Tobias Salz for sharing his slides on the topic!

Beliefs

- Today's lecture builds on the previous lecture on belief updating
- Will consider one of the applications that I consider most promising: how people update their beliefs when interacting with AI systems
- Start with deep dive into "Combining Human Expertise with Artificial Intelligence Experimental Evidence from Radiology" (Agarwal et al. 2023)
  - By Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz
- Currently R&R at Econometrica

Introduction

Artificial Intelligence (AI) is a transformative technology

- General purpose (Brynjolfsson & Mitchell 2017, Brynjolfsson et al. 2017, Lai et al. 2021)
- Targeted at high-skilled tasks relative to previous technologies (Webb 2020)

AI tools have outperformed humans in predictive domains

- Examples include credit scoring, bail judgements, medical diagnostics ... (Kleinberg et al. 2018, Liu et al. 2019)

Radiology is an iconic example:

"We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists"

— Geoffrey Hinton (in 2016)
(Obermeyer & Emanuel 2016)

Will AI Replace Radiologists?

"The right answer is: Radiologists who use AI will replace radiologists who don't."

— Curtis Langlotz (2019)

"Human in the loop" is the common presumption in the medical profession:

1. Legal and regulatory hurdles
2. Partial task automation
3. Correct mistakes made by the AI

Are human expertise and predictive AI complements or substitutes?

- Should they be combined? How and why?

  ... setting aside legal/regulatory barriers

Research Questions

Humans relative strengths and weaknesses

- ✓ Have access to valuable (non-systematic) information
- X May not combine sources of information effectively
- X Positive marginal cost of time

Questions:

1. Is contextual information valuable?
2. How do humans combine AI predictions with their own information?
3. How should we design human-AI collaboration?

Research Approach

1. Conduct an **information experiment** with radiologists
   - Retrospective cases of potential chest pathologies

   Key features
   - Varies the availability of AI assistance and contextual information
   - Measure diagnostic assessments and treatment/follow-up recommendations

2. Compare **posterior beliefs** relative to a Bayesian benchmark. Find:
   i. Automation neglect (Dietvorst et al. 2015)
   ii. Correlation/signal dependence neglect (Enke & Zimmermann 2019)

3. In paper: **Design** the collaboration between humans and machines
   – Trade-off: Human time cost vs diagnostic quality
   – AI signal is always available for free

Related Literature

Human versus AI prediction in medicine
Rajpurkar et al., 2017; Irvin et al., 2019; Mullainathan & Obermeyer, 2019; Ribers & Ullrich, 2022

Human–AI collaboration in radiology
Patel et al., 2019; Rajpurkar et al., 2022; Seah et al., 2021; Fogliato et al., 2022

Delegation / learning to defer
Mozannar & Sontag, 2020; Bansal et al., 2021

Biased belief updating
Grether, 1980, 1992; Benjamin, 2019; Enke & Zimmermann, 2019; Conlon et al., 2022

Economics of AI and human use of AI tools
Agrawal et al., 2018, 2019; Kleinberg et al., 2017; Angelova et al., 2022; Noy & Zhang, 2023

This paper's niche: measures behavioral biases in a naturalistic expert setting and links them to optimal human–AI deployment.

Outline

Classification Problem

Actions, state and payoffs

- Correct decision $\omega_i \in \{0, 1\}$ with prior probabilities $\pi(\omega)$
- Expert decision-maker r takes a binary action $a_{ir} \in \{0, 1\}$
- Payoff

$$u(a_{ir}, \omega_i) = -\mathbb{1}\{a_{ir} = 1, \omega_i = 0\} \cdot c_{FP} - \mathbb{1}\{a_{ir} = 0, \omega_i = 1\} \cdot c_{FN}$$

Signal distribution $\pi\left(s_{A,i}, s_{H,ir} \,|\omega_i\right)$

- $s_{H,ir}$: expert's signal
- $s_{A,i}$: AI signal

Classification Problem

Optimal action, given $s \subseteq \{s_{A,i}, s_{H,ir}\}$, defined by **cut-off rule**:

$$a_r^*(s_{ir}) = \mathbb{1} \cdot \left\{ \frac{p_r(\omega_i = 1|s_{ir})}{p_r(\omega_i = 0|s_{ir})} > c_{rel} \equiv \frac{c_{FP}}{c_{FN}} \right\},$$

where posterior beliefs may be incorrect

$$p_r(\omega_i|s_{A,i}, s_{H,ir}) \overset{?}{\neq} \pi(\omega_i|s_{A,i}, s_{H,ir})$$

Two key quantities of interest

1. Relative costs of false negatives/positives c_{rel}
   - **Estimated** using follow-up recommendations
2. Prior and updates odds-ratios with and without AI
   - **Elicited** via the experiment

Outline

Overview of the Experimental Design

2 x 2 (x 2) Design

Treatment Dimension 1: Access to AI prediction (AI)

Treatment Dimension 2: Clinical History (CH)

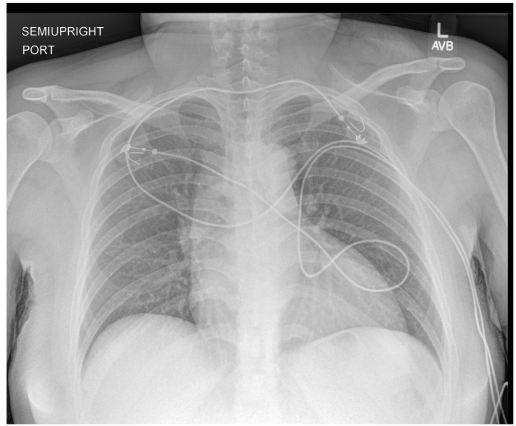
(Treatment Dimension 3: Incentives for Accuracy (Hossain & Okui 2013))

Radiologists **participate remotely** through tailormade interface

- Mimics clinical practice but generates structured quantifiable report
- In collaboration with radiologists at Stanford and Mt. Sinai (NYC)
- 324 historical cases from Stanford Healthcare System with Chest-X-ray and clinical history, manually reviewed for public release

✓ Structured data entry v. free text report

X-ray Landing Page



Zoom In | Zoom Out | Reset X-Ray | Full Screen

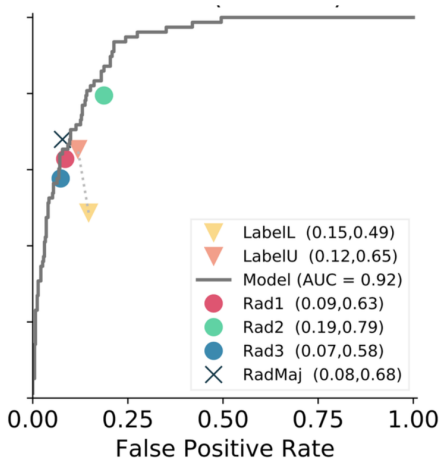
Contrast

Brightness

Treatment Dimension 1: AI Algorithm

CheXperT: new iteration of CheXNet

- Large convolutional neural net
- Trained on $\geq 200,000$ chest X-rays
- Probabilities for 14 pathologies
- Uses only the X-ray image, not clinical history or vitals
- Ground-truth used in training based on past reports
- Performance superior or similar to board-certified radiologists



→ **AI treatments** give access to CheXperT's probability for likelihood of disease.

Treatment Dimension 2: Clinical History

Contextual information provided

- Treating physician indication / clinical note
- Patient vitals
- Laboratory results

Why this matters

- Valuable information available to radiologists in practice
- Harder to incorporate into AI training because of logistical and privacy constraints

Indication

30 years of age, Female, history of hypertension, abnormal EKG, abdominal pain, evaluate for cardiomegaly or mediastinal widening.

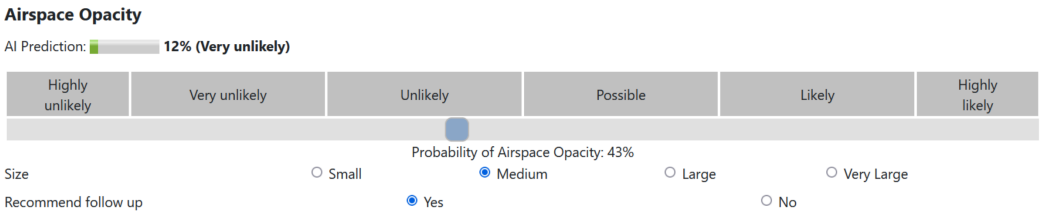
Vitals

Variable	Value
Weight	170 lbs
BP	243/166 mmHg
Temp	99.1F
Pulse	99.0 bpm
Age	30

Abnormal Labs All Labs

Variable	Value	Unit	Flag
ALT (SGPT), Ser/Plas	38.0	U/L	High
AST (SGOT), Ser/Plas	39.0	U/L	High
Eosinophil, Absolute	0.01	K/uL	Low

Assessments and Decisions



- ▶ Elicit assessments in a hierarchy of pathologies to minimize burden
- ✓ Analyze pre-registered groups
  1. All pathologies
  2. All pathologies with AI predictions
  3. Top-level pathologies with AI predictions
  4. Overall abnormal/normal assessment

Diagnostic Standard

Diagnostic standard ω_i constructed using aggregate assessment of experts

- Five board certified chest radiologists from Mount Sinai Health Care System
- Follows the medical AI literature (Irvin et al. 2019, McCluskey et al. 2021)

Definitive diagnostic test typically unavailable

- Many thoracic pathologies do not have clean non-imaging-based ground truth
- Selective labels problem when administered (Mullainathan & Obermeyer 2022)

Baseline uses cutoff at $\bar{p} = 0.5$ (Wallsten & Diederich 2001)

- Robust to log-odds averaging
- Robustness to comparisons with $\bar{p}$

Subject Recruitment and Training

Radiologists are well suited experimental subjects

- Remote work is common $\rightarrow$ No interaction with patients
- Tele-radiologists hired on spot market (receive a piece-rate $\approx$ \$10 per read)

Subject training

- Instructions with a demo-video
- Explanation of AI tool, comprehension questions
- 50 example patients showing the X-ray and AI output
- Comprehension questions before the study begins

Data

- 227 tele-radiology participants
- $60 - 240$ case reads each

Experimental Design

Challenges:

- Expensive subject pool $\rightarrow$ Power
- Construct Bayesian benchmark:

$$\ln \frac{\pi(\omega_i = 1 | s_{A,i}, s_{H,ir})}{\pi(\omega_i = 0 | s_{A,i}, s_{H,ir})} = \ln \frac{\pi(s_{A,i} | \omega_i = 1, s_{H,ir})}{\pi(s_{A,i} | \omega_i = 0, s_{H,ir})} + \ln \frac{\pi(\omega_i = 1 | s_{H,ir})}{\pi(\omega_i = 0 | s_{H,ir})}$$

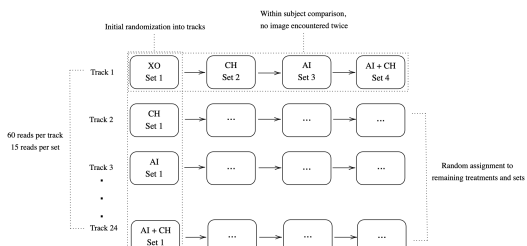
- Linked assessments within radiologist-case are ideal
  1. Elicit beliefs without AI assessment and construct Bayesian benchmark
  2. Compare with posterior beliefs elicited with AI assessment

Approach: Hybrid design that collects both within and across subject data

Design 1

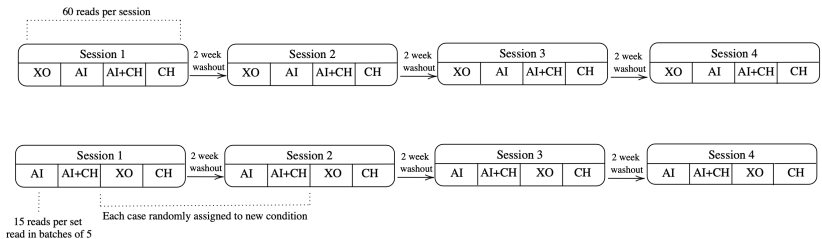
Simple across design with a within subject component

- ✓ Clear across design
- ✓ Within subject comparison hedges power
- X Comparing with Bayesian benchmark not possible
- X Across component only powered for large effect sizes
- ▶ 159 tele-radiologists, 60 cases each



Design 2

- ✓ Wash-out period to get a fresh read under different modality
- ✓ Allows comparing with Bayesian benchmark
- ▶ 240 reads for each radiologist



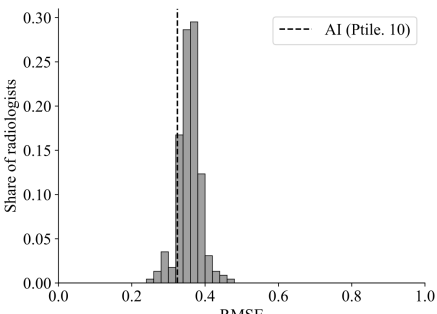
Concern: Subjects may remember previously provided information

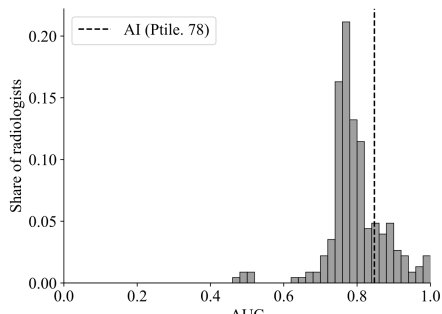
- ▶ No movement towards AI in subsequent reads
- ▶ Design 3 ($\sim$50 rads, 100 cases): reads w/ AI predictions only after reads w/o AI

AI Performance

Radiologists and AI performance:

- Top level pathologies with AI prediction
- Algorithm performs better than most radiologists





Outline

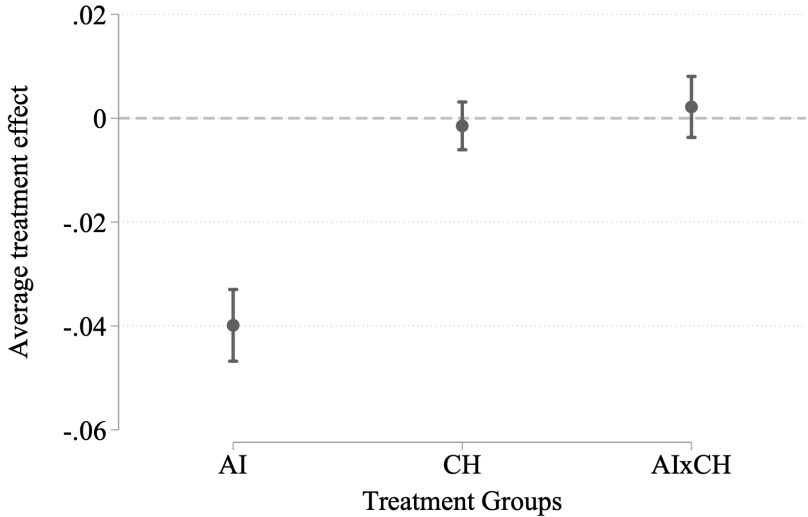
Specification

Main specification for treatment-effect analysis:

$$Y_{irt} = \gamma_{g_i} + \gamma_{CH}\, d_{CH}(t) + \gamma_{AI}\, d_{AI}(t) + \gamma_{AI \times CH}\, d_{CH}(t) d_{AI}(t) + \varepsilon_{irt}$$

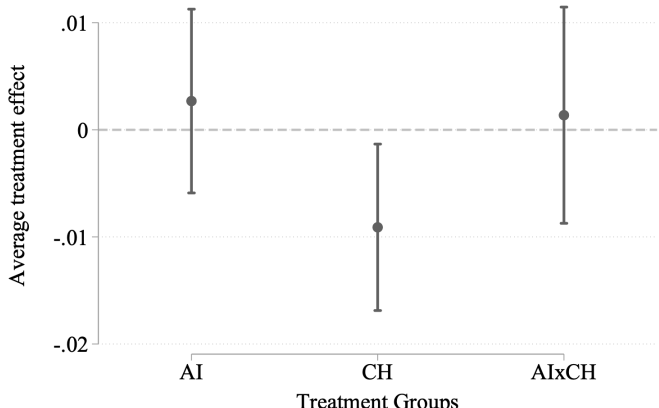
- γ_{g_i}: pathology fixed effects
- Two-way clustered standard errors at the radiologist and patient level
- Main outcomes: deviation from diagnostic standard, deviation from AI, incorrect decision, and active time

Treatment Effect — Deviation from AI



Treatment Effect — Deviation from Diagnostic Standard

- AI assistance has no significant average effect on performance
- Clinical history improves average diagnostic performance
- The AI $\times$ CH interaction is also close to zero



Conditional Effect of AI Assistance

- Average effects mask systematic heterogeneity
- AI helps when radiologists are uncertain
- AI also helps when the AI is very confident that a pathology is absent
- AI hurts when the AI signal is uncertain
- AI also hurts when, absent AI, radiologists are very confident that no pathology is present

$\rightarrow$ This pattern rejects Bayesian updating with correct beliefs.

Outline

Incorrect Beliefs/Updating

Describe updating via building on Grether, 1980, 1992; Benjamin, 2019:

$$\log \frac{p(\omega_i = 1|s_{A,i}, s_{H,ir})}{p(\omega_i = 0|s_{A,i}, s_{H,ir})} = b \log \frac{\pi(s_{A,i}|\omega_i = 1, \tilde{s})}{\pi(s_{A,i}|\omega_i = 0, \tilde{s})} + d \log \frac{\pi(\omega_i = 1|s_{H,ir})}{\pi(\omega_i = 0|s_{H,ir})} + \varepsilon_{ir}$$

Terminology:

- Automation bias / neglect: $b \gtrless d$
- Signal dependence neglect: the AI update term behaves as if it does not condition on the radiologist's own signal $\tilde{s} = \{\emptyset, s_{H,ir}\}$

Theoretical implication:

- If only automation neglect is present, AI weakly improves decisions
- Signal dependence neglect can make AI reduce performance for some signals

Specification and Estimation Challenges

Empirical analog:

$$\log \frac{P_{ir}^{AI}}{1 - P_{ir}^{AI}} = a + b \cdot llr(s_{A,i}; \tilde{s}) + d \cdot lor(s_{H,ir}) + \varepsilon_{ir}$$

where P_{ir}^{AI} is the reported probability with AI assistance.

Main challenges:

1. Need the same radiologist-case read with and without AI assistance
2. The AI update term is a ratio of conditional densities
3. The radiologist's own signal s_H is not directly observed
4. Measurement error in elicited beliefs can attenuate estimates

Solutions

Rewrite the AI update term using Bayes' rule:

$$llr(s_{A,i}; \tilde{s}) = \log \frac{\pi(\omega_i = 1 | s_{A,i}, \tilde{s})}{\pi(\omega_i = 0 | s_{A,i}, \tilde{s})} - \log \frac{\pi(\omega_i = 1 | \tilde{s})}{\pi(\omega_i = 0 | \tilde{s})}$$

1. Use radiologist reports without AI to proxy for own information
2. Estimate the conditional term flexibly using pathology-specific random forests
3. Allow richer controls using reported probabilities on other pathologies
4. Use reports from other radiologists as instruments to address measurement error
5. Estimate the updating equation with two-step GMM

How Do Radiologists Update Their Beliefs?

Radiologists may have incorrect beliefs about signal structure, for example:

1. Act "as if" AI and own signals are conditionally independent
2. Ignore information in other pathologies

Empirical strategy:

- Estimate the updating equation using designs 2 and 3
- Compare non-nested models that differ in the update term
- Bayesian updating with correct beliefs corresponds to $b = d = 1$ and Constant $= 0$
- Use J-statistics and Rivers–Vuong tests for model comparison

Evidence of Automation and Signal Dependence Neglect

- Selected model: radiologists behave as if AI and own signals are conditionally independent
- Cross-pathology information does not improve fit in the main top-level-pathology specification
- Estimated coefficients in the selected model:
  - AI term: $b = 0.26$
  - Own-information term: $d = 0.87$
- Interpretation:
  - Automation neglect: radiologists underweight AI relative to their own information
  - Signal dependence neglect: radiologists do not properly account for overlap between AI and own signals

$\rightarrow$ These two biases can make AI reduce performance for some signals.

Model Selection: J-Statistic and Rivers–Vuong Tests

- The latest paper compares the candidate updating models using GMM-based specification tests, not BIC
- J-statistic:
  - Model (1): 1.62
  - Model (2): 29.38
  - Model (3): 28.39
- Pairwise Rivers–Vuong tests reject models (2) and (3) relative to model (1)
- Main takeaway: the preferred model combines automation neglect with signal dependence neglect

Designing Human–AI Collaboration

- The paper solves an optimal delegation problem over three modalities:
  - AI only
  - Human only
  - Human + AI
- Even when AI improves human decisions, AI alone often performs better than Human+AI
- Human time costs are similar with and without AI assistance
- For low false-negative costs, AI decides almost all cases
- As false-negative costs rise, more cases are delegated to humans, but mostly without AI assistance

→ In the observed data, cases are typically assigned to AI alone or human alone, rarely to Human+AI.

Conclusion

Objective: understand how experts use AI and how collaboration should be designed

Main findings:

1. AI assistance does not improve average diagnostic quality, but contextual information does
2. Radiologists move toward the AI, yet combine information imperfectly
3. The best-fitting model implies automation neglect and signal dependence neglect
4. AI helps for some signals but hurts when the AI is uncertain
5. Optimal delegation usually assigns cases to AI alone or human alone, rarely to Human+AI

Economists can contribute to the design of human–AI collaboration

- Information design and delegation, organizational incentives, training and learning

References I

Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. (2023), Combining human expertise with artificial intelligence: Experimental evidence from radiology, NBER Working Paper 31422, National Bureau of Economic Research.

Brynjolfsson, E. & Mitchell, T. (2017), 'What can machine learning do? workforce implications', *Science* **358**(6370), 1530–1534.

Brynjolfsson, E., Rock, D. & Syverson, C. (2017), Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics, NBER Working Paper 24001, National Bureau of Economic Research.

Dietvorst, B. J., Simmons, J. P. & Massey, C. (2015), 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General* **144**(1), 114–126.

Enke, B. & Zimmermann, F. (2019), 'Correlation neglect in belief formation', *The Review of Economic Studies* **86**(1), 313–332.

Hossain, T. & Okui, R. (2013), 'The binarized scoring rule', *The Review of Economic Studies* **80**(3), 984–1001.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K. et al. (2019), CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 33, pp. 590–597.

References II

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. (2018), 'Human decisions and machine predictions', *The Quarterly Journal of Economics* **133**(1), 237–293.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A. & Tan, C. (2021), 'Towards a science of human-AI decision making: A survey of empirical studies', *arXiv preprint arXiv:2112.11471* .

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A. & Denniston, A. K. (2019), 'A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis', *The Lancet Digital Health* **1**(6), e271–e297.

McCluskey, R., Enshaei, A. & Hasan, B. A. S. (2021), 'Finding the ground-truth from multiple labellers: Why parameters of the task matter', *arXiv preprint arXiv:2102.08482* .

Mullainathan, S. & Obermeyer, Z. (2022), 'Diagnosing physician error: A machine learning approach to low-value health care', *The Quarterly Journal of Economics* **137**(2), 679–727.

Obermeyer, Z. & Emanuel, E. J. (2016), 'Predicting the future — big data, machine learning, and clinical medicine', *The New England Journal of Medicine* **375**(13), 1216–1219.

References III

Wallsten, T. S. & Diederich, A. (2001), 'Understanding pooled subjective probability estimates', *Mathematical Social Sciences* **41**(1), 1–18.

Webb, M. (2020), The impact of artificial intelligence on the labor market, Working paper, Stanford University.