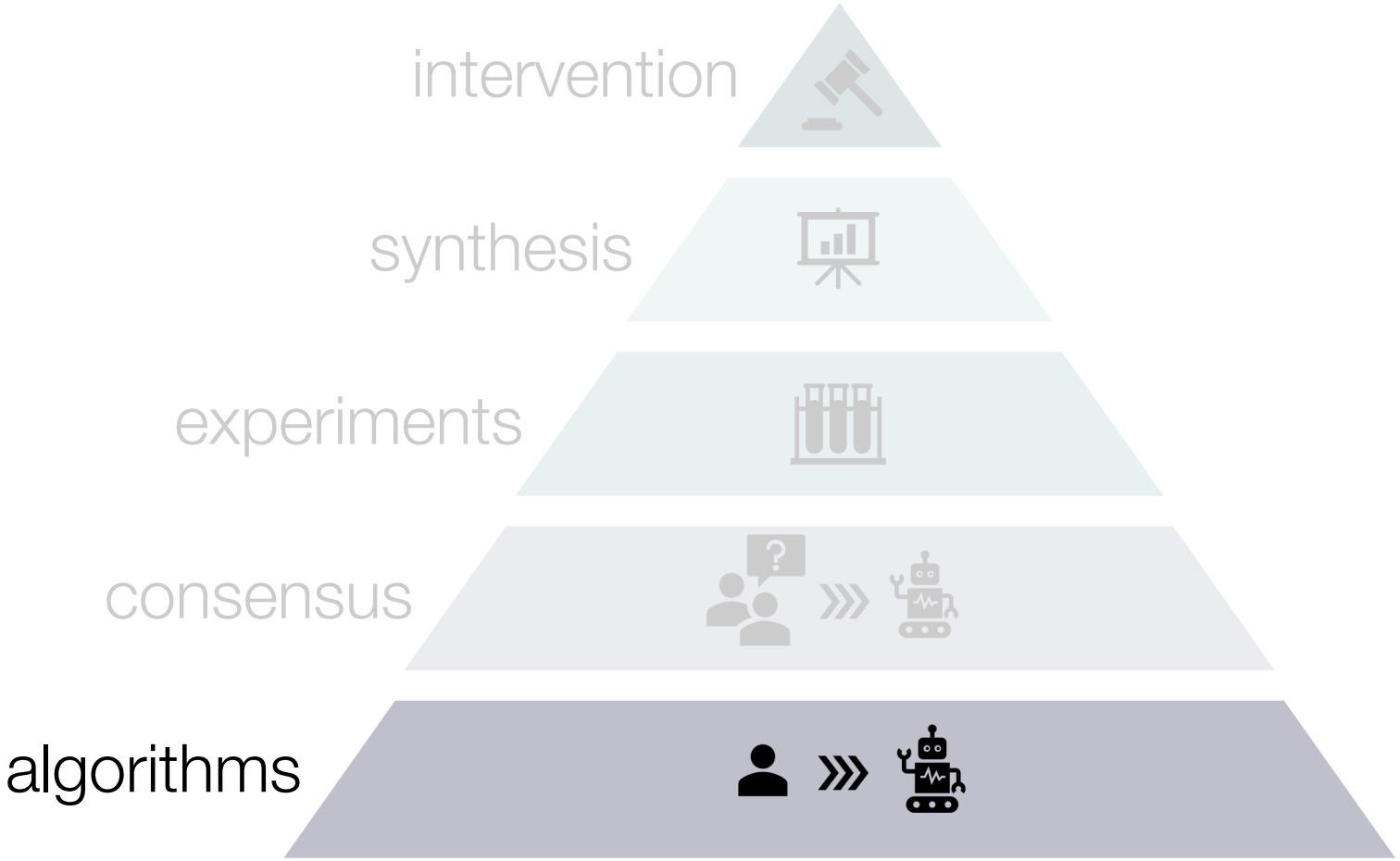
Evidence-based decision making

The problems of algorithms

Rui Mata, FS 2025

Version: March 17, 2024

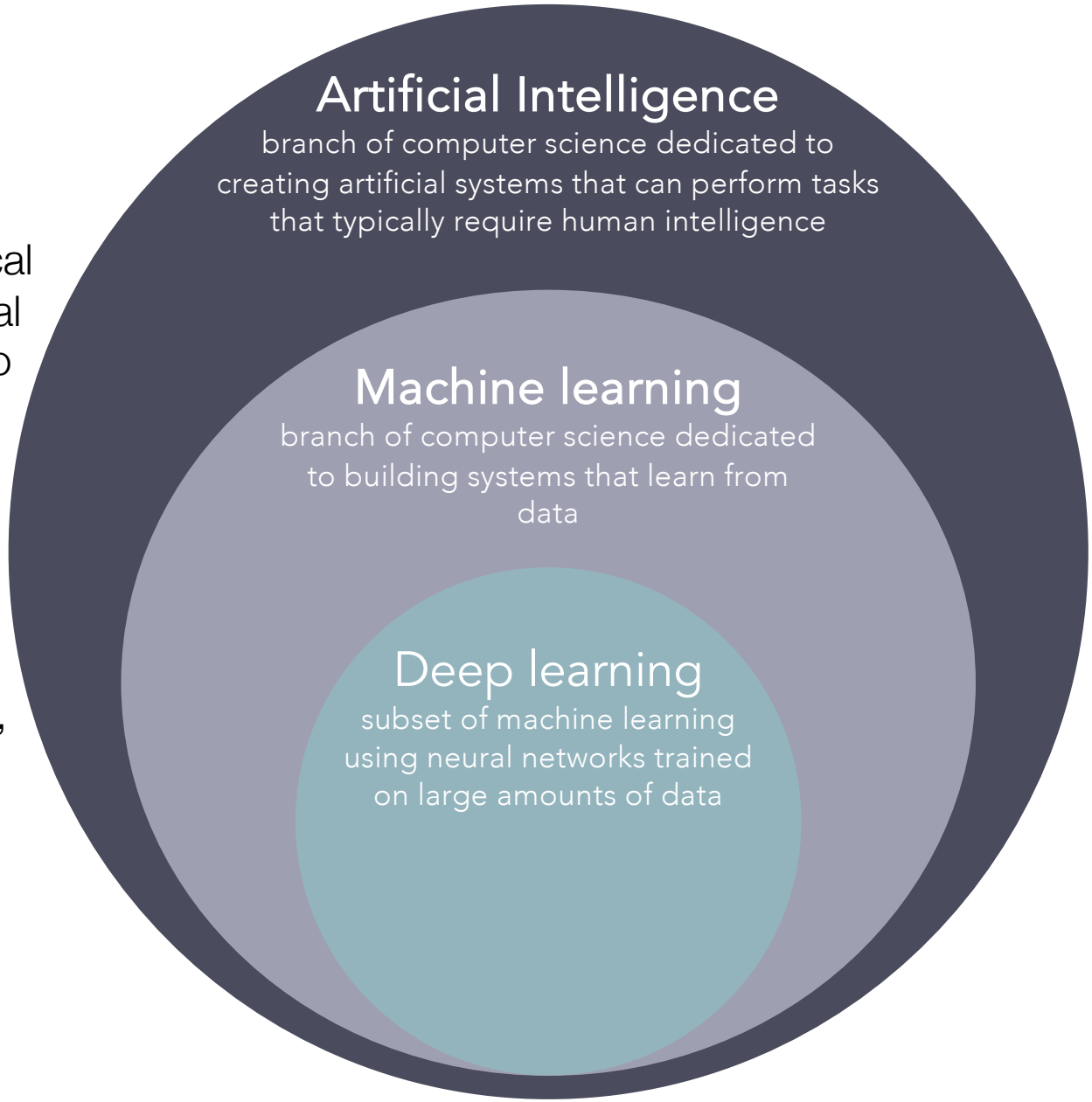

intervention
synthesis
experiments
consensus
algorithms

Goals for today

- Discuss the relation between actuarial judgment and definitions of artificial intelligence, machine learning, and deep learning
- Discuss problems of **overfitting** when using actuarial approaches and possible solutions
- Discuss the problem of **bias** in actuarial approaches and possible solutions
- Discuss issues concerning the **adoption of/trust in** actuarial judgment in practice

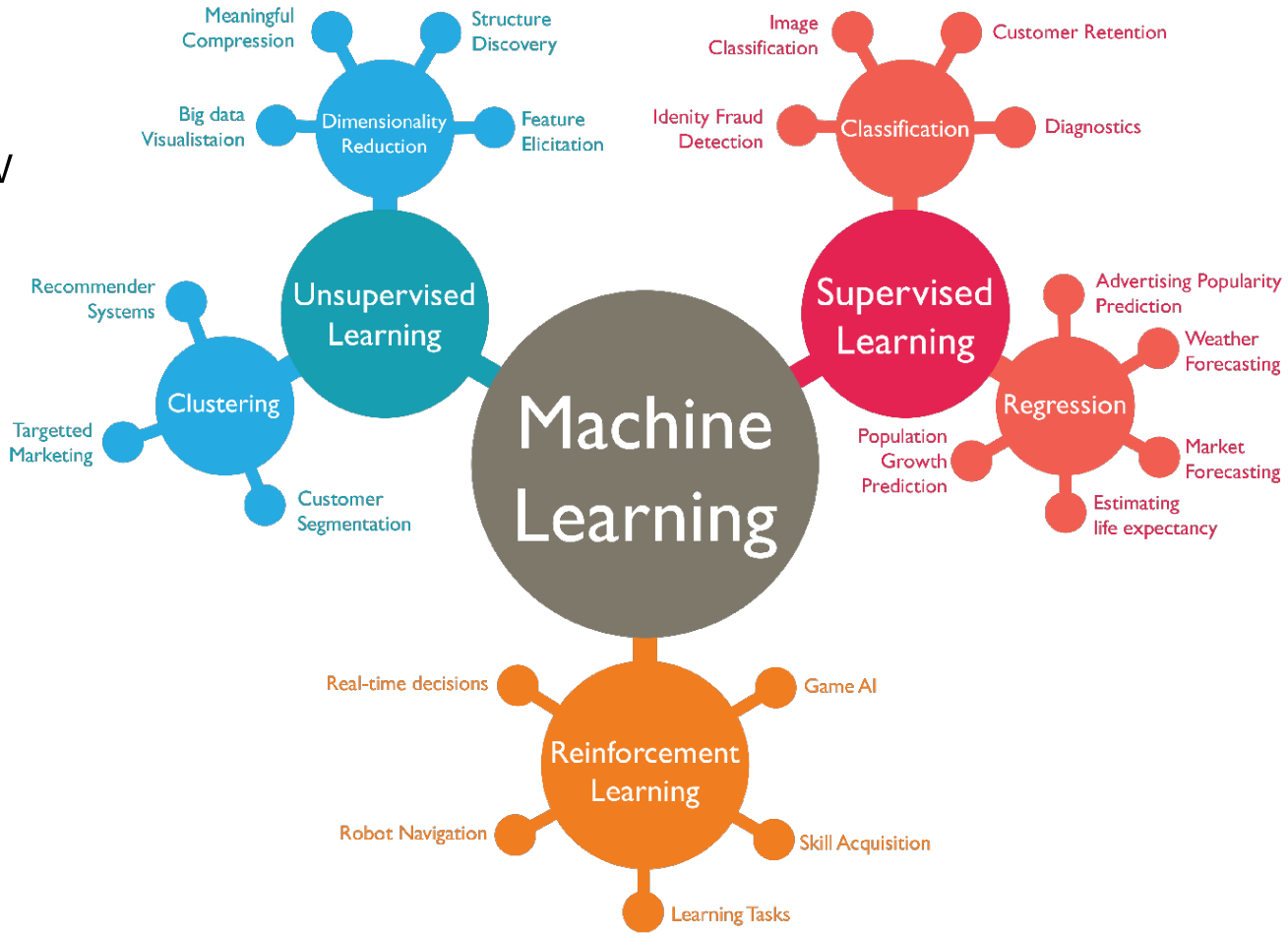
Actuarial judgment

refers to decision-making processes that rely on statistical models, algorithms, or actuarial tables. These tools are used to predict outcomes based on quantitative data from past cases. Actuarial judgment emphasizes objectivity and consistency, using empirical data and predefined rules to make predictions or decisions, without the influence of the decision maker's momentary state.



Different types of machine learning…

"Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?"



Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. http://doi.org/10.1126/science.aaa8415

Bias-variance dilemma

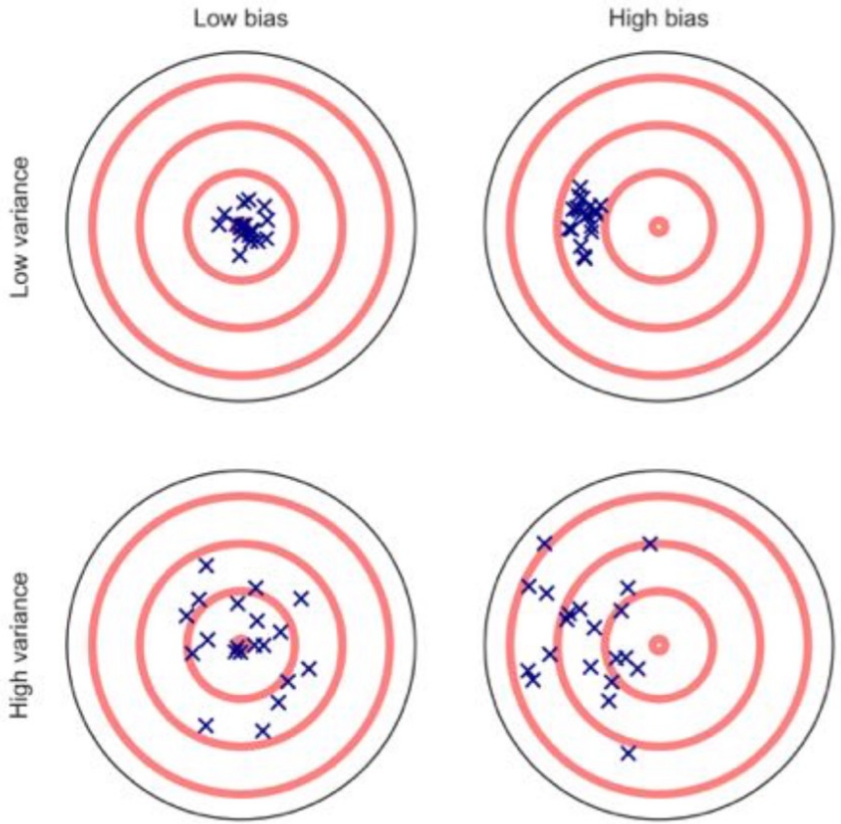


Figure 2. An estimator's predictions can deviate from the desired outcome (or true scores) in two ways. First, the predictions may display a systematic tendency (or *bias*) to deviate from the central tendency of the true scores (compare right panels with left panels). Second, the predictions may show a high degree of *variance*, or imprecision (compare bottom panels with top panels).

A dilemma exists because bias and variance are not independent: Methods for reducing variance tend to increase bias, and methods for reducing bias tend to increase variance (Brighton & Gigerenzer, 2015)

Yarkoni, T. & Westafall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research, 68*(8), 1772–1784. http://doi.org/10.1016/j.jbusres.2015.01.061

Overfitting

When a model learns patterns specific to the training data but fails to generalize to new data.

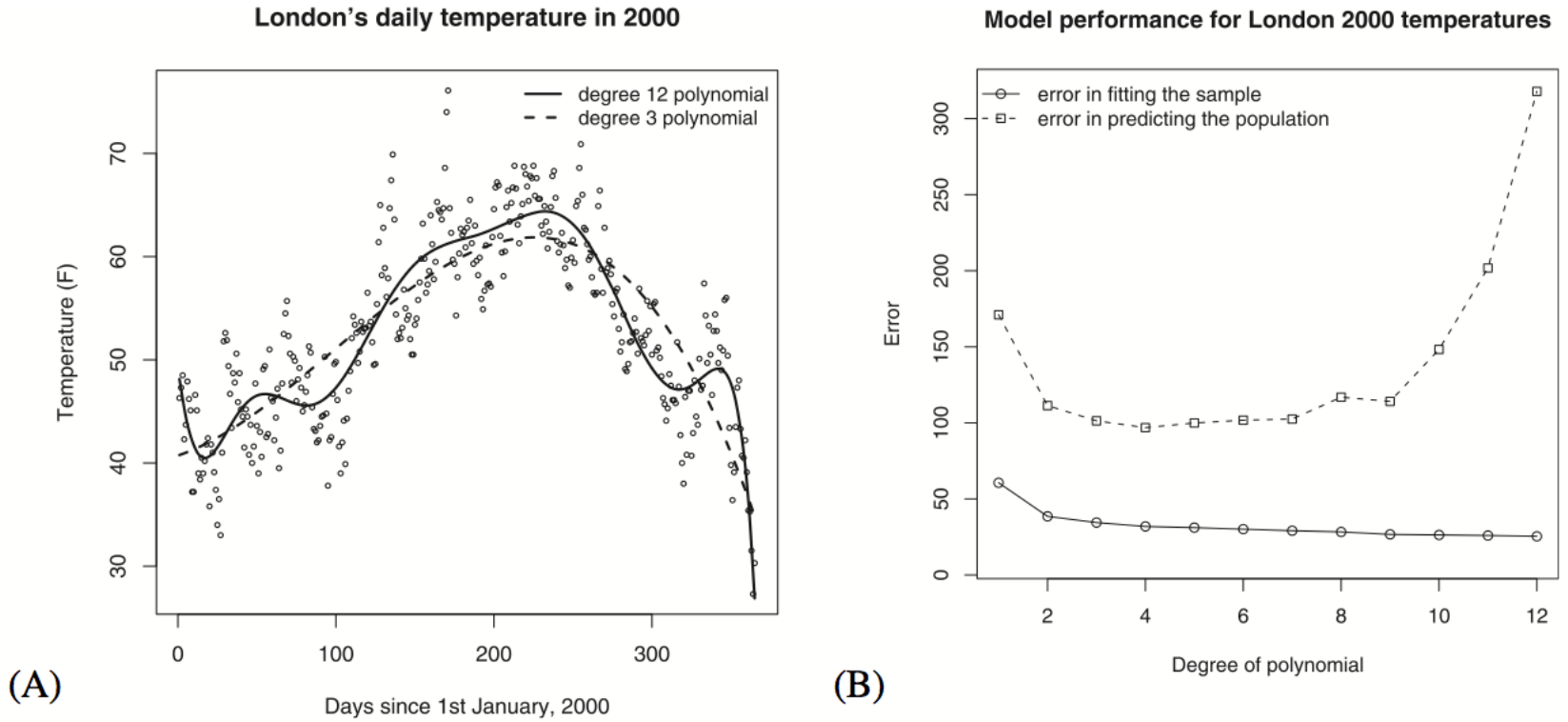


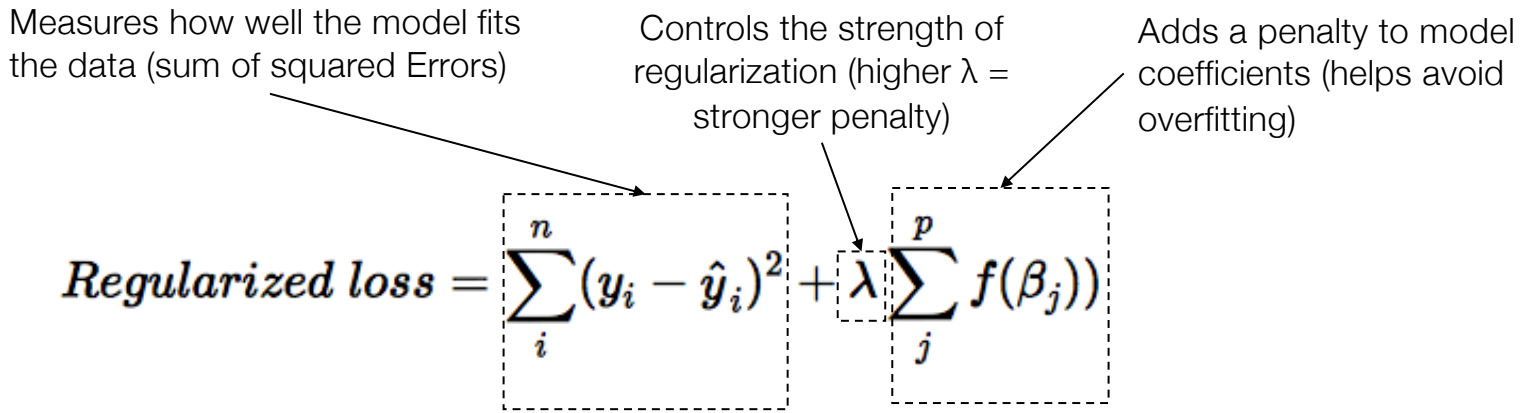
Fig. 3. Plot (A) shows London's mean daily temperature in 2000, along with two polynomial models fitted with using the least squares method. The first is a degree-3 polynomial, and the second is a degree-12 polynomial. Plot (B) shows the mean error in fitting samples of 30 observations and the mean prediction error of the same models, both as a function of degree of polynomial.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why biased minds make better inferences. ***Topics in Cognitive Science, 1(1)***, 107–143. doi:10.1111/j.1756-8765.2008.01006.x

but see Dubova et al. (2025). Is Ockham's razor losing its edge? New perspectives on the principle of model parsimony. Proceedings of the National Academy of Sciences, 122(5), e2401230121. https://doi.org/10.1073/pnas.2401230121

How can one avoid overfitting?

- **Apply regularization**: A technique to prevent overfitting by adding a penalty to coefficients in a model.



$f(\beta_j)$ can take different forms:

lasso regression: |β| (βs are reduced in size, resulting in automatic feature selection, with some βs becoming zero)

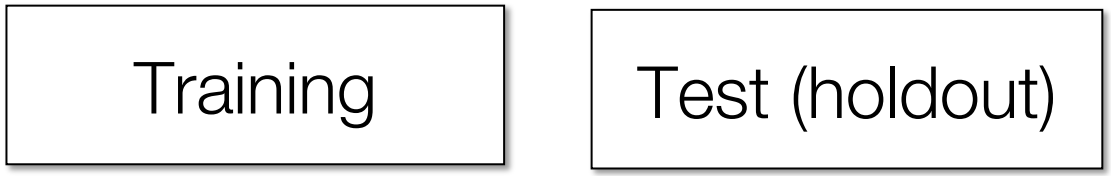
ridge regression: β^2 (squaring reduces the size of extreme βs).

elastic net: |β| + β^2 (a combination of both)

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. ***Perspectives on Psychological Science, 12***(6), 1100-1122.

How can one avoid overfitting?

- **Cross-validation**: Compare models in how well they predict out-of-sample



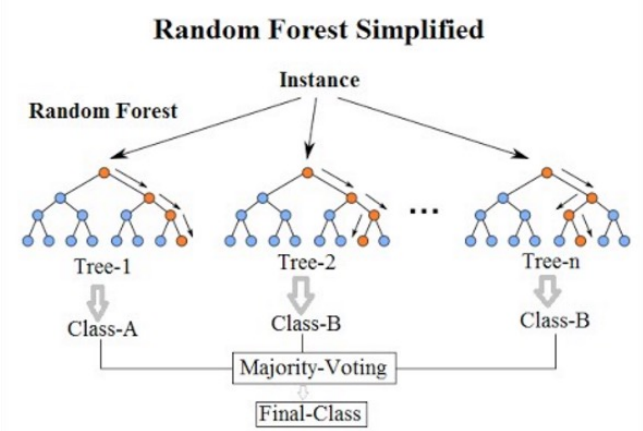
Cross-validation is a technique used to evaluate the performance of machine learning models by splitting the dataset into multiple subsets, helping to prevent overfitting and improve generalization. **Holdout Validation** is a simple approach that splits the data into separate training and test sets (figure above). The most common approach, however, is ***K*-Fold Cross-Validation**, where the data is divided into K equal parts, and the model is trained and tested K times, each time using a different fold as the test set. A number of other variations exist that have different computational requirements and advantages/disadvantages.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. ***Perspectives on Psychological Science, 12***(6), 1100-1122.

How can one avoid overfitting?

- **Averaging**: modeling approaches that integrate averaging or use different models and combine their predictions (ensembles), for example, random forests

A decision tree is a flowchart-like model used for classification and regression, where data is split into branches based on feature conditions until a decision is reached at the leaf nodes. It is easy to interpret but prone to overfitting. **Random Forests** improves upon this by creating an ensemble of multiple decision trees, each trained on random subsets of the data and features. It then combines their predictions—using majority voting for classification or averaging for regression—resulting in a more robust, accurate, and less overfitting-prone model.



Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. Annual Review of Economics, 11(1), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

O’Neil

“algorithms are opinions embedded in code!”

Some critique:

False Objectivity – People tend to assume that algorithms are neutral and objective because they are based on math. However, they are shaped by human decisions about what data to use and how to weigh different factors.

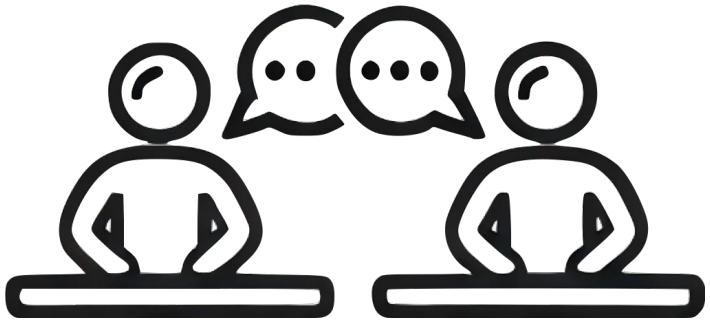
Opaque and Unaccountable – Many algorithms operate as "black boxes," meaning their inner workings are not transparent. This makes it difficult to challenge biased decisions.

Use of Flawed or Biased Data – If historical data used to train algorithms contains biases, the algorithm will replicate and even amplify those biases. For example, hiring algorithms trained on past successful employees may discriminate against underrepresented groups.

Self-Reinforcing Feedback Loops – Algorithms can reinforce existing biases. For example, predictive policing algorithms might send more officers to neighborhoods with higher reported crime rates, leading to more arrests and confirming the initial bias.

WHAT POTENTIAL SOURCES OF BIAS ARE THERE?

Discuss how to represent different sources of bias using the lens model framework - consider the example of using digital tracing to uncover worker's personality



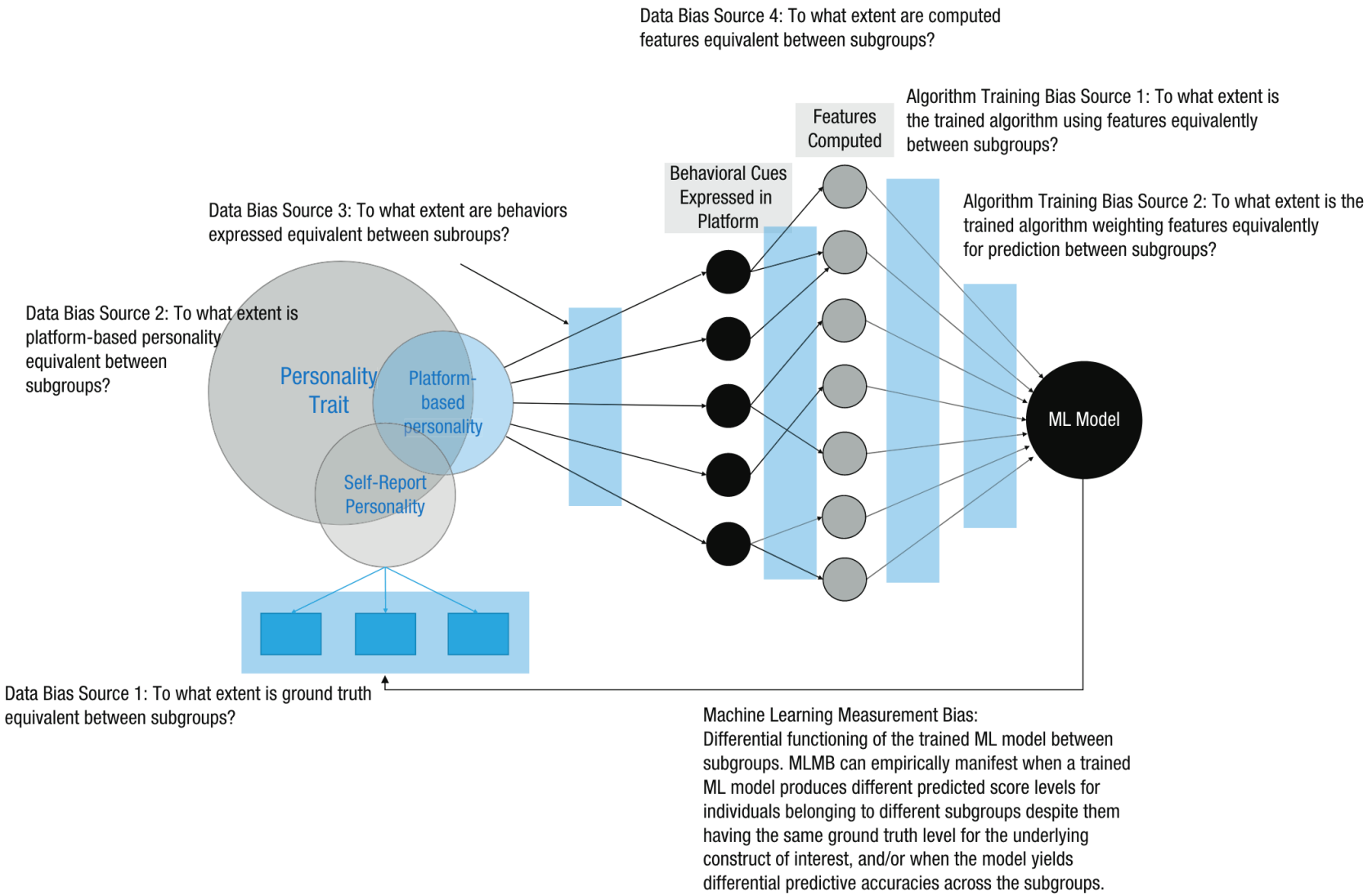


Fig. 4. Expanding the Brunswik lens model to identify the sources of machine-learning measurement bias: an illustration using personality as the focal construct. Areas highlighted in blue represent possible sources of machine-learning measurement bias; "platform-based personality": the personality construct measured by input data (e.g., online personality assessed by social media data) used in machine-learning models to predict self-report personality.

Assessing Bias

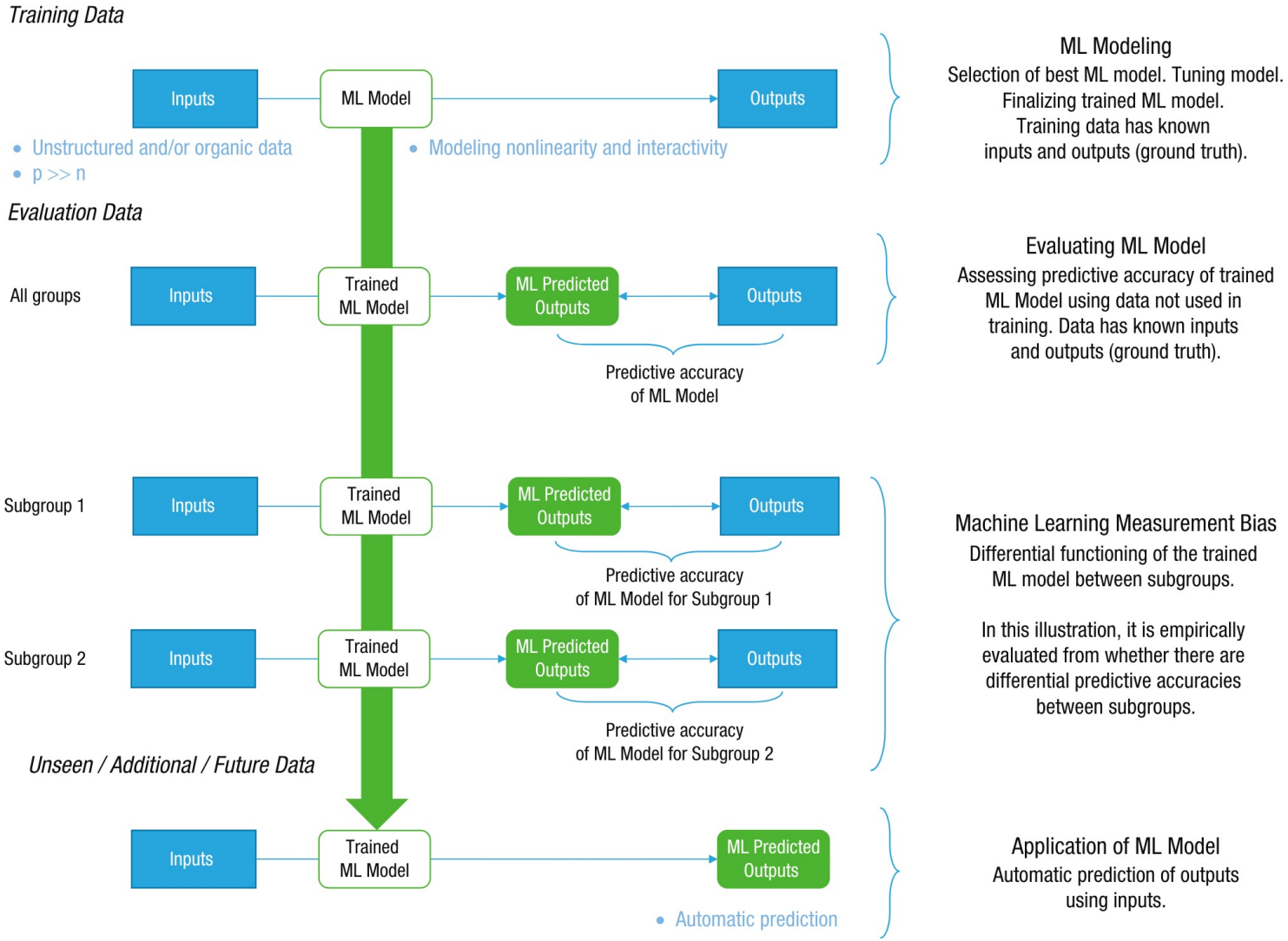


Fig. 1. Simplified process of machine-learning modeling.

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science, 5* (1),1–30.

Table 2. Comparison Between Traditional Measurement Bias and Machine-Learning Measurement Bias

Key issues	Measurement bias	Machine-learning measurement bias
Types of scores that are relevant	Predicted observed scores typically derived from CFA or IRT models of psychological assessments Latent scores typically derived from CFA or IRT models of psychological assessments	ML-model-predicted scores that are predictions produced by the ML model Ground-truth scores typically in the form of observed scores from psychological assessments
Defining bias	Defined as a differential relationship between the latent score and the predicted observed score or differential functioning of the measurement tool across subgroups One empirical manifestation is that the measurement model produces different scores for individuals belonging to different subgroups despite the same latent-score level. Another empirical manifestation is that the same measurement model does not fit subgroups equally well.	Defined as differential functioning of the trained ML model between subgroups One empirical manifestation is when a trained ML model produces different predicted score levels for individuals belonging to different subgroups despite them having the same ground-truth level for the underlying construct of interest. Another empirical manifestation is that the ML model yields differential predictive accuracies across the subgroups.
Empirical manifestation of bias	Most typically assessed via differences in model-data fit: (a) differences in CFA fit between subgroups and (b) item-level subgroup differences in IRT fit Can also be assessed based on different model-predicted scores for the same latent-trait level	Ground-truth score level: different ML-predicted score levels between subgroups when subgroups have the same ground-truth score level Ground-truth distribution level: different ML-predicted score distributions (e.g., means, variances) between subgroups for equivalent subgroup ground-truth distributions or the discrepancy between ML-predicted subgroup score distributions and ground-truth subgroup score distributions Predictive accuracy: different ML-model prediction accuracies (i.e., nonequivalent convergence of predicted scores and ground-truth scores) between subgroups Modeling ground-truth score and ML-predicted scores: applying (regression) models between ground-truth scores and ML-predicted scores and finding that significantly different models are needed between subgroups

Note: CFA = confirmatory factor analysis; IRT = item response theory; ML = machine learning.

Adoption of actuarial methods

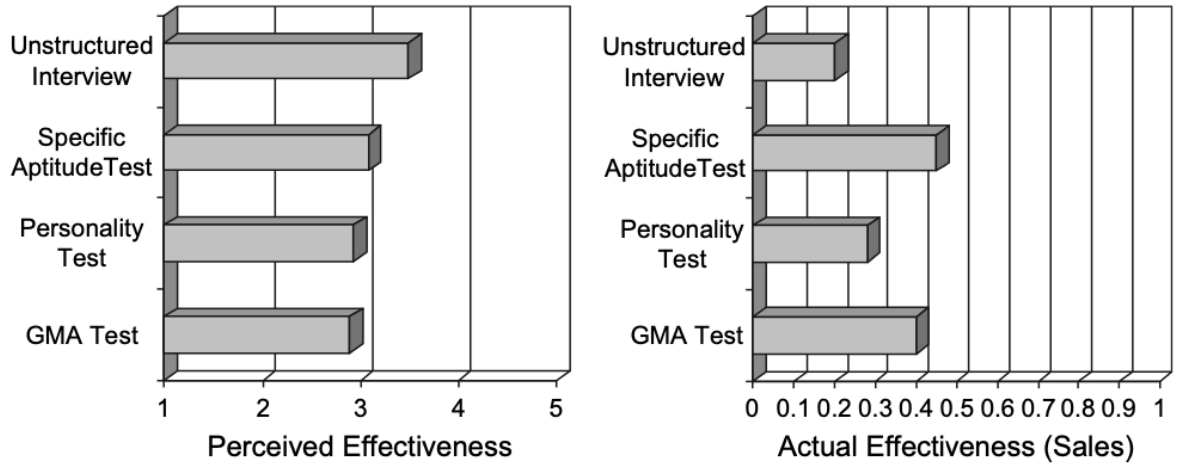


Figure 1. Perceived versus actual usefulness of various predictors.
Note. Perceived effectiveness numbers are on a 1–5 scale (1 = *not good*; 3 = *average*; 5 = *extremely good*). Actual effectiveness numbers are correlations corrected for unreliability in the criterion and range restriction. Because Vinchur, Schippmann, Switzer, and Roth (1998) did not include interviews, the interview estimate is from Huffcutt and Arthur (1994) level 1 interview. GMA = general mental ability; personality = potency; specific aptitude = sales ability.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1*(3), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Adoption of actuarial methods: Algorithm aversion

Dietvorst et al. conducted experiments where participants were shown the performance of both algorithms and humans on prediction tasks. The participants were then asked to choose whether they would rely on the algorithm or a human for future predictions, sometimes after seeing one or both make errors. The results showed the tendency of people to avoid using algorithms after they have seen them make mistakes, even if the algorithms have a better overall performance than humans

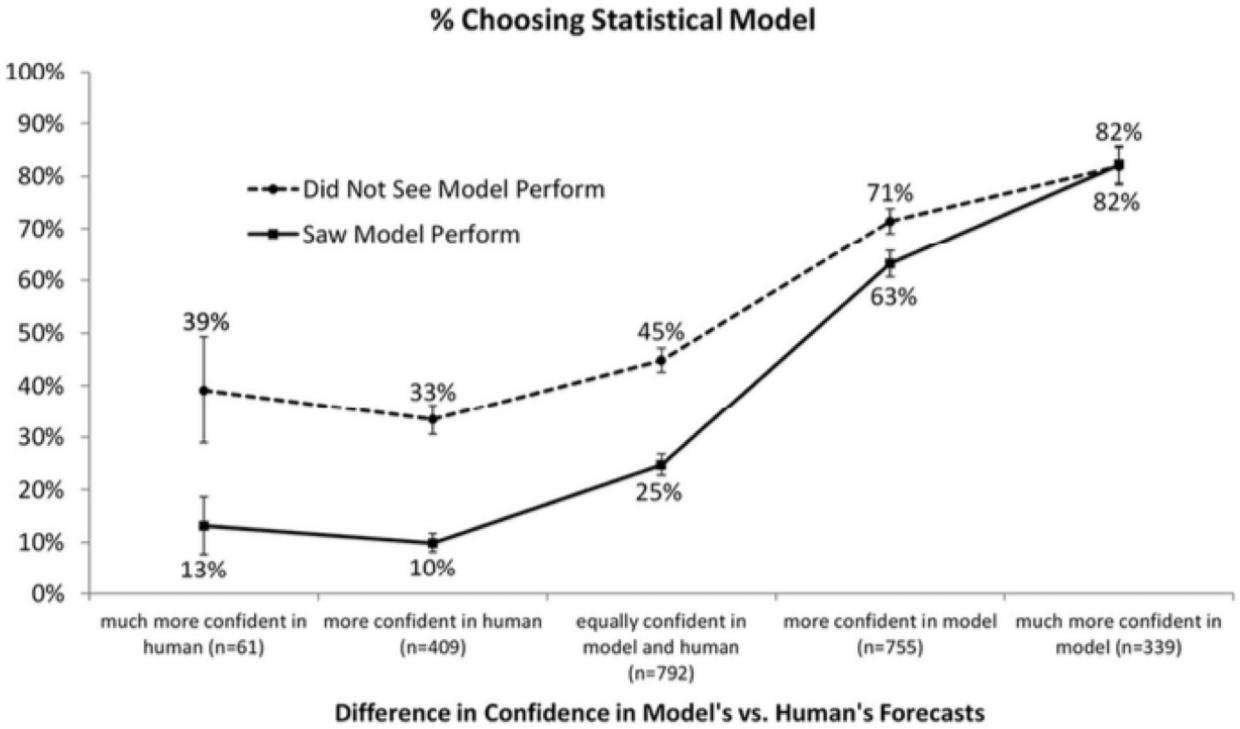


Figure 4. Most people do not choose the statistical model unless they are more confident in the model's forecasts than in the human's forecasts. Errors bars indicate ±1 standard error. The "Did Not See Model Perform" line represents results from participants in the control and human conditions. The "Saw Model Perform" line represents results from participants in the model and model-and-human conditions. Differences in confidence between the model's and human's forecasts were computed by subtracting participants' ratings of confidence in the human forecasts from their ratings of confidence in the model's forecasts (i.e., by subtracting one 5-point scale from the other). From left to right, the five *x*-axis categories reflect difference scores of: <–1, −1, 0, +1, and >1. The figure includes results from all five studies.

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. http://doi.org/10.1037/xge0000033

Adoption of actuarial methods: Algorithm aversion

False expectations -> decision makers may have incorrect/unrealistic beliefs about the performance of algorithms

Lack of decision control -> decision makers may strive for autonomy

Lack of incentivization -> incentives for algorithmic use may be unclear or misaligned (effort vs. performance)

Combatting intuition -> decision makers may have incorrect (overconfident) beliefs about own intuition

Conflicting concepts of rationality -> lack of a match between algorithm's knowledge and those of the individual (risk vs. uncertainty)

Burton, J. W., Stein, M. K., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. ***Journal of Behavioral Decision Making***, ***27***(11), 1309–20. http://doi.org/10.1002/bdm.2155

Adoption of actuarial methods

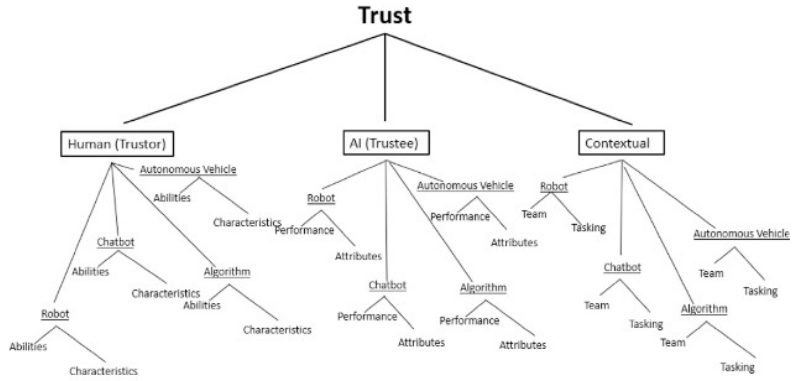


TABLE 2: Human-Related Antecedents of Trust in AI

Antecedent of Trust	K	Effect Size *D*	S^2_g	S^2_e	95% CI Upper	95% CI Lower
Global	23	0.26*	0.34	0.09	0.14	0.38
Ability-based	8	0.32*	0.60	0.02	0.21	0.42
Competency/Understanding	2	1.02*	0.05	0.18	0.43	1.61
Expectancy	8	0.29	0.63	0.33	−.11	0.69
Expertise	4	0.47*	0.14	0.20	0.03	0.91
Operator performance	2	0.76	0.02	1.46	−.92	2.43
Prior experience	4	−.19	0.69	0.06	−.43	0.05
Workload	2	−1.19	0.23	1.38	−2.82	0.44
Characteristic-based	20	0.38*	0.18	0.11	0.24	0.53
Age	2	0.09	0.02	0.02	−.08	0.26
Attitudes toward AI	5	1.05	0.30	3.61	−.61	2.72
Comfort with AI	1	−.37	---	---	---	---
Culture	2	0.51*	0.04	0.07	0.15	0.87
Education	1	0.17	---	---	---	---
Gender	3	0.42*	0.05	0.05	0.17	0.67
Personality traits	4	0.25*	0.47	0.02	0.12	0.37
Propensity to trust	1	0.70	---	---	---	---
Satisfaction	1	1.04	---	---	---	---

Note. *Denotes significance at the $p < .05$ level. ** s^2_e = sampling error variance; s^2_g = observed variance.

TABLE 3: AI-Related Antecedents of Trust

Antecedent of Trust	K	Effect Size *D*	S^2_g	S^2_e	95% CI Upper	95% CI Lower
Global	48	0.62*	1.10	0.09	0.54	0.70
Performance-based	22	1.47*	1.34	0.17	1.30	1.64
Dependability	2	0.80	0.15	2.02	−1.18	2.77
Performance	13	1.48*	1.41	0.16	1.26	1.70
Predictability	2	1.42	0.67	1.85	−.46	3.31
Reliability	5	2.70*	0.33	0.37	2.16	3.23
Attribute-based	35	0.31*	0.55	0.07	0.22	0.39
AI Personality	4	0.63*	2.39	0.04	0.42	0.83
Anthropomorphism	10	0.30*	0.29	0.12	0.08	0.52
Appearance	1	−.05			---	---
Behavior	6	0.81*	0.38	0.09	0.57	1.04
Communication	9	0.06	0.15	0.05	−.08	0.20
Level of automation	2	0.03	0.00	0.01	−.10	0.17
Reputation	5	0.68*	0.04	0.12	0.38	0.99
Transparency	9	0.24*	0.26	0.06	0.08	0.40

Note. *Denotes significance at the $p < .05$ level. ** s^2_e = sampling error variance; s^2_g = observed variance.

Trust in AI depends both on human and AI characteristics

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors: The Journal of the Human Factors and Ergonomics Society, 65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Summary

- **Actuarial judgement:** Actuarial judgment has direct links to machine learning techniques because it often implies using statistical models that learn from data to perform estimation or categorization tasks.
- **Overfitting:** Overfitting occurs when models' predictions are tuned to noise rather than signal in available data and more complex models are not always better because such models may be affected by noise; machine learning techniques offer potential remedies, including regularization, cross-validation, and ensemble methods.
- **Bias:** A problem emerging from the use of actuarial approaches is the codifying of undesired "opinions" in code; conceptual frameworks exist for investigating and mitigating bias in machine learning applications, typically involving some form of auditing models for specific biases.
- **Algorithm adoption**: actuarial approaches are not always applied in practice; there is an ongoing academic debate centered around the reasons for lack of adoption of algorithms in professional settings, including the role of training, expectations, incentives, etc.; more work is needed…