A TAXONOMY OF HUMAN AND ML STRENGTHSIN DECISION-MAKING: ACLASSIFICATION

K. Sujatha[1] V.M. Venkateswara Rao[2]

[1,2]Associate Professor.
[1,2]Shree Ramchandra College of Engineering, Pune, Maharashtra.
[1]dsv0414@gmail.com, [2]vsrao.ee@gmail.com

Abstract:

All along the machine learning (ML) pipeline, people are essential. An intricate network of dispersed assessments allows people to annotate unmatched volumes of data, which aids in the development of machine learning systems. Expert practitioners work with ML model outcomes in a range of real-world areas, including healthcare, lending, education, social services, and disaster relief, on the machine learning deployment end. In order to better integrate human judgment with machine learning algorithms, this paper looks at and supports it in complicated decision-making scenarios. Building on the rich and fertile ground from human behavior-studiesing disciplines, such as psychology, cognitive science, and human-computer interaction, this work studies the situated human factors in various socio-technical systems, like crowdsourcing, peer review, and ML-assisted decision-making, from both quantitative and qualitative perspectives. To be more precise, we create statistical instruments to comprehend human behaviour in various data elicitation scenarios. Next, in order to assist evidence-based policy reform aimed at improving decision quality, we design experiments that yield statistically sound insights regarding human decision-making biases in complicated environments. We propose both domain-specific and domain-general frameworks to facilitate efficient human-ML collaboration, with the goal of enhancing ML deployment in real-world scenarios. Understanding and utilizing the relative advantages of people and machine learning methods is the main goal here. This study demonstrates the value of highlighting human involvement in the larger endeavour to increase the effectiveness of machine learning algorithms.

1. Introduction:

The use of machine learning (ML) models in decision-making systems has increased dramatically in recent years, with applications found in a wide range of industries, including criminal justice, credit lending, and healthcare (1), for example. In the criminal justice system, algorithmic recidivism risk scores inform pre-trial bail decisions for defendants (4), and in credit lending, lenders regularly use credit-scoring models to assess the risk of default by applicants (5). The excitement surrounding the use of these technologies in high-stakes decision-making is fuelled by the promise of these technologies to tap into large datasets, mine relevant statistical patterns within them, and use those patterns to make more accurate predictions at a lower cost and without experiencing the same cognitive bias.

However, a growing body of research indicates that ML models are susceptible to a number of biases (6) and instability (7). Furthermore, because machines lack human characteristics like cognitive flexibility, social and contextual knowledge, and commonsense reasoning abilities, they frequently result in detrimental outcomes

in practice (8). Due to these findings, there have been requests for both human and machine learning participation in high-stakes decision-making systems. The idea behind these requests is to create carefully crafted hybrid decision-making systems that combine and enhance the advantages of both ML models and human thought processes. In actuality, such systems are widespread, even in the previously listed sectors. A variety of hybrid human-ML designs, from algorithm-in-the-loop (10) to human in-the-loop (9) configurations, have been developed and tested by researchers.

Nevertheless, there are conflicting empirical results about these solutions' efficacy and success (11). Concurrently, an expanding corpus of theoretical research has endeavoured to formulate and structure these hybrid frameworks (12) and examine the most effective methods for combining human and machine learning assessments inside them (12).We examined the literature on human behaviour, cognitive and behavioural sciences, and psychology to develop our taxonomy of human-ML complementarity and identify the key areas where human and machine learning decision-making processes diverge. utilizing the traditions of cognitive science, computational social science, behaviour, cognitive and behavioural sciences, and psychology to comprehend the fundamental differences between the decision-making processes of humans and machines. As per the conventions of cognitive science and computational social science (13), we comprehend decision-making in both human and machine learning.

Our taxonomy delineates specific distinctions between human and machine learning decision-making. We propose a mathematical framework that encompasses all the factors in the taxonomy and shows how our taxonomy can be used to analyse whether we can expect complementarity in a given environment and what kinds of human-ML combination can help accomplish it. Specifically, we formulate an optimization issue for convex combinations of decisions made by humans and machine learning. This issue formulation creates a framework for academics to investigate which traits of humans and machine learning models can support complementing performance. We suggest quantitative complementarity measures in order to classify various forms of complementarity. These measures are intended to capture two prominent forms of human-ML collaboration that have been identified in the literature: communication-based collaboration and routing (or deferral).

We simulate ideal human-ML pairings under two different scenarios to illustrate the application of our taxonomy, the optimization problem setup, and the related metrics of complementarity: (1) Different feature sets are available to human and machine learning models; (2) distinct objective functions are associated with each model type. Through the comparison of optimal aggregation procedures under these circumstances, we are able to obtain important insights into how each agent contributes to the best possible combined choice. This helps future research and practice create human-ML relationships in these circumstances in a successful manner. Together, these studies demonstrate that integrating human-ML judgments should take use of each entity's distinct advantages and disadvantages, since various sources of complementarity influence the kind and degree of performance improvement that may be attained through human-ML collaboration.

2. Methodology for designing the taxonomy

In order to look into the possibility of complementarity in combined human-ML decision-making, we must comprehend the advantages and disadvantages of the ML model and the human decision-maker in the particular application. As an example, it has been noted that ML models make conclusions based on far larger data sets than humans could effectively process (13). Human decision-makers may be unable to replicate the rich contextual knowledge and common-sense reasoning abilities that humans bring to the decision-making process (10). As a result, we create a taxonomy for human-ML decision-making that takes into consideration

the many distinctions between machine learning and human decision-makers, including applications that involve predictive decision-making.

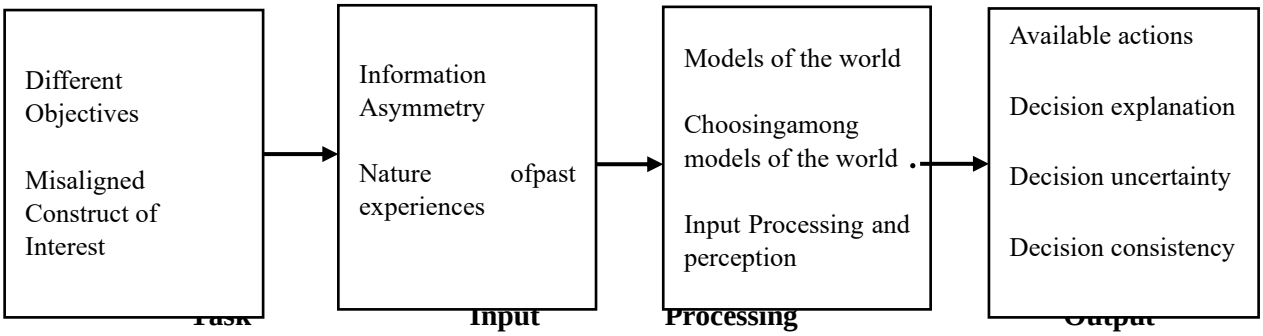

Figure 1: Proposed taxonomy of human and ML strengths & weaknesses in decision-making divided into four parts of the decision-making process: Task definition,
input, internal processing, and output.

3. A taxonomy of human and ML strengths & weaknesses in decision-making

In this study, we analyse two decision-making agents: the ML model, represented by M, and the human, represented by H. Using the notation we expand upon it by using$X_{H,}$ $X_M$to represent the feature space that each agent has access to, where $X_{H,}$ $X_M \subseteq$ X. Similarly, we consider a human version and an ML version for each variable presented for our decision-making setup in the previous section, denoted by subscripts H and M, respectively.

3.1. Task

The key differences between the ML model and the human in terms of how the decision-making task is defined.
• **Objective:** For example, supervised learning models seek to reduce anticipated loss while reinforcement learning models want to maximize expected cumulative rewards. These are the sole predicted performance optimization goals of most machine learning models. Although recent studies have investigated methods for developing models that take into account a wider range of goals, such as various risk metrics (15), definitions of fairness (16), and ideas of interpretability (17). encoding every facet of the goals that a human decision-maker would seek to maximize is either challenging or impractical (18). Our notation allows us to write this as $F_H \neq F_M$. For instance, while deciding whether to lend money, bankers may take into account a variety of risk variables as well as other criteria including upholding their organization's lending policies and customer connections (19)
• **Misaligned construct of interest:** When machine learning is used in social situations, theoretical factors like socioeconomic status, teacher efficacy, and recidivism riskall of which are difficult to quantify are frequently included in the models. Rather, they are deduced indirectly using proxies, which are measures of attributes that are noted in the data that a model may access.

Simplifying assumptions must be made in order to define proxy variables for a construct of interest, and there is frequently a conceptual gap between ML proxies and how human decision-makers understand the targeted construct (20). Stated otherwise, $O_H(X, a) \neq O_M(X, a)$. According to Jacobs and Wallach (21), the mismatch between the notion of interests and the inferred measures is the direct cause of a number of harms examined

in the literature on the fairness of socio-technical systems. Obermeyer et al. (22), for instance, looked at racial biases in a hospital-based machine learning platform. They discovered that poorer healthcare provision decisions were made for Black patients compared to White patients when patients' healthcare expenditures were used as an indirect proxy to predict patients' demand for treatment. In this case, the proxythe monetary cost of treatmentwas easily obtained from the data that was at hand, but it is very different from how medical practitioners understand patients' true needs for care.

3.2. Input

We now go over the unique features found in the inputs that ML models and humans employ.

Access to different information:
From an input standpoint, both people and robots have access to shared and non-overlapping information in many contexts, such as criminal justice and healthcare: $X_H \neq X_M$. This is because important traits that cannot be codified for machine learning are frequently present in real-world decision-making scenarios. For instance, a physician can examine a patient's physical presentation and have a better understanding of their symptoms since this data is difficult to encode and give to the computer. In a similar vein, interactions with the defendant teach the judge about their predispositions (4). In the research on human-ML complementarily, this phenomenon is also known as unobservable (10) and information asymmetry (23).

Nature of past experiences:
The training datasets utilized by contemporary machine learning systems are significantly different from the nature of embodied human experience over the course of a lifetime: $D_H \neq D_M$. For instance, ML models are frequently trained on a large number of historical examples of a particular decision-making activity, yet the training data comprises a confined and unchanging set of data. This frequently fails to capture the depth of human experience. People make judgments based on a lifetime of experiences from a variety of fields, and it can be challenging to identify the precise facts that they consider. In the other hand, people usually only learn from large amounts of training data, whereas machine learning models can learn from small portions of a large number of human decision-makers' judgments.

3.3. Internal processing

The key differences between the internal workings of ML systems and humans are now discussed.

Models of the world: Humans rely on sophisticated mental models and "theories" that store complicated views about causal mechanisms in the universe, not merely statistical associations, as is thoroughly reviewed in(24). As a result, human representations of the world differ from those represented by machine learning models (ML models):$\pi_H \neq \pi_M$. For instance, beginning Humans acquire complex belief systems about the social and physical worlds from a young age (intuitive psychology and intuitive physics), which greatly influence how they see and interpret their surroundings. Human mental models are often compositional and causal, in contrast to contemporary machine learning methods.Consequently, humans can learn more quickly than contemporary machine learning systems thanks to these strong preconceived notions about the world, and they can draw conclusions from very little amounts of data such as one-shot and few-shot learning (25) Conversely, nevertheless, the model. Whether it is a class of parametric or non-parametric models, the machine decision-maker class has a more mathematically tractable form (26). While researchers frequently use data and model

architecture to encode domain knowledge when designing these models, like neural networks, most machine learning models still have issues with distribution shift (27), lack of interpretability (28), and require large sample sizes.

Choosing among models of the world:

ML models look for the model that maximizes their goal differently than humans do based on the task specification, data, and models of the environment $OPT_H \neq OPT_M$. Due to their vastness, modern machine learning models such neural networks can need a significant amount of processing power and are often learnt by first-order approaches. according to the models (29). Conversely, heuristics that humans can use in a very short length of time are possible (30). These straightforward tactics could be superior than more intricate models in situations when there is a lot of inherent uncertainty in the work. We direct readers to (31) for a more thorough analysis of the situations in which and situations in which such heuristics would be more appropriate.

3.4. Output

We now go over the unique qualities that set human and machine learning system outputs apart.

Available actions

The range of options or actions that ML models and humans can choose from in real-world deployment circumstances can vary: $A_H \neq A_M$. For instance, in the context of K–12 education, ML-based tutoring software would be able to provide kids just the right amount of guidance when they're having trouble with a math concept. In the meanwhile, even though a humanAlthough teachers using this software in the classroom have less time to spend with each student, they can still support students in more ways, like by offering emotional support or assisting with prerequisite material that is not covered by the software (10).Similar to this, a model may only be able to suggest that a case be looked into or not, depending on the data that is currently available, in the context of ML-assisted child abuse screening. On the other hand(32) human call screeners may take steps to get further information as necessary, such as calling other parties who may be relevant to a case.

Explaining the decision

The ability of humans and machine learning to explain the reasoning behind their choices varies. Extensive research has been conducted on interpretability and explain ability (XAI) for machine learning (33). people are typically more adept than machine learning algorithms in producing coherent explanations that have significance for other people, according to research in cognitive and social psychology. Additionally, (34) contends that XAI The focus of study should shift from vague, subjective ideas of what constitutes a "good" explanation to the rationale and decision-making processes that individuals use to choose an explanation. They discover that human explanations are, above all, social and contextual, contrastive, and biased in their selection. However, human explanations could not match their true underlying decision-making processes (35) ,In contrast, we can always track the exact computational processes that resulted in the output prediction when using machine learning models (36).

Uncertainty communication:

New techniques for calibrating an ML model's prediction uncertainty have been developed in response to growing research in the field of uncertainty quantification for machine learning (37). Additionally, techniques have been developed to break down the uncertainty in the model into two categories: epistemic uncertainty,

which represents the inherent randomness in an application domain and cannot be reduced, and epistemic uncertainty, also referred to as systematic uncertainty, which represents the uncertainty resulting from a lack of knowledge or information and can be reduced (38). These techniques for quantifying uncertainty, however, might not always be well-calibrated (37)and are a current line of inquiry. In the meanwhile, human decision-makers typically produce discrete decisions rather than uncertainty scores and struggle to calibrate their level of uncertainty or confidence in their choices (39). Additionally, Individuals differ in their uncertainty calibration scales (40).

Output consistency:

When a decision-maker consistently generates the same result for the same input, we say that they are consistent. As a result, we take into account the inconsistent judgments that stem from extraneous elements, which we refer to as factors unrelated to the input. The time of day, the weather, and other variables are a few instances of external elements. Studies in Human behavior and psychology have demonstrated the inconsistent nature of human judgments (41). More precisely, given the exact identical issue description at two separate times, there is a positive chance that a particular human decision-maker would arrive at a different choice. Numerous fields, including clinical psychology (43), medicine (42), finance, and management (41), have shown within-person inconsistencies in human judgments. This type of discrepancy is not shown using common ML algorithms.2

Time efficiency:
ML models can provide more judgments in a shorter amount of time than human decision-makers in a variety of scenarios. Humans frequently have very little time for decision-making overall, in addition to perhaps taking longer for each decision.

4. Conclusion

Our research advances our knowledge of potential processes behind complementing performance in human-machine learning. By combining knowledge from several different fields of study, we offer a taxonomy that describes the possible advantages of human and machine learning-based decision-making. Our taxonomy offers a platform for academics and professionals working on human-ML collaboration to examine and comprehend the possible causes for anticipating cooperative group performance within the relevant application fields. With the use of this taxonomy, we believe the research community will be able to articulate its expectations regarding the contexts in which they anticipate complementarity between human and machine learning for decision-making.
We provide a problem design for the best convex combination of human and machine learning judgments, along with related measures for complementarity, based on our taxonomy. Our suggested framework integrates a number of earlier methods for merging judgments made by humans with machine learning. Crucially, a critical study of our approach indicates that the best way to integrate human and machine learning-based decisions will rely on the particular relative advantages each has in the particular decision-making application area. As the simulations in Section 3 show, our optimization approach may be used to create hypotheses regarding the best methods to combine human and ML-based assessments in specific scenarios.
Models of human and machine decision-making, as well as historical decision-making data, can be used for this.

By contrasting the different outcomes of these simulations, academics and practitioners may also better grasp

the trade-offs associated with integrating human-ML collaboration in a decision-making scenario. increases in accuracy compared to implementation costs. It is important to keep in mind that, although the joint decision-maker is an idealized form in theory, actual joint decision-making may have lesser accuracy owing to human decision-making's inefficiencies. Therefore, before implementing, it would be helpful to estimate the possible advantages of collaborative decision-making. It would also be extremely beneficial from a theoretical and practical one to examine the hypotheses and trade-offs that our models offer empirically.

References:

1. Bhavik N. Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A. J. Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew Lungren. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digital Medicine, 2(1), December 2019. ISSN 2398-6352.
2. Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable machine learning in credit risk management. Computational Economics, 57, 01 2021.
3. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, machine-bias-risk-assessments-in-criminal-sentencing, 2016.
4. Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 469–481, 2020.
5. Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 5125–5131, 2013. ISSN 0957-4174.
6. Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296, 2018.
7. Ali Alkhatib. To live in their utopia: Why algorithmic systems create absurd outcomes. In Proceedings of the 2021 CHI conference on human factors in computing systems, pages 1–9, 2021.
8. Olga Russakovsky, Li-Jia Li, and Fei-Fei Li. Best of both worlds: Human-machine collaboration for object annotation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition, pages 2121–2131, 2015.
9. Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the loop: Decisions in the presence of erroneous algorithmic scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.
10. Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. Toward supporting perceptual complementarity in human-ai collaboration via reflection on unobservables. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1):1–20, 2023.
11. Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. Human-ai collaboration with bandit feedback. arXiv preprint arXiv:2105.10614, 2021.
12. David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In Advances in Neural Information Processing Systems, pages 6147–6157, 2018.
13. Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and brain sciences, 40, 2017.

14. Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. Business Horizons, 61, 07 2018.
15. Justin Khim, Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. Uniform convergence of rank weighted learning. In International Conference on Machine Learning, pages 5254–5263.
16. Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5):82–89, 2020.
17. Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. Proc. ACM Hum.-Comput. Interact., 5(CSCW2), oct 2021.
18. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. The quarterly journal of economics, 133(1):237– 293, 2018.
19. Carl-Christian Tronnberg and Sven Hemlin. Lending decision making in banks: A critical incident study of loan officers. European Management Journal, 32(2):362–372, 2014.
20. Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–33, 2021.
21. Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 375–385, 2021.
22. Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453, 2019
23. Patrick Hemmer, Max Schemmer, Niklas Kuhl, Michael V ¨ ossing, and Gerhard Satzger. On the ¨ effect of information asymmetry in human-ai teams. arXiv preprint arXiv:2205.01467, 2022.
24. Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and brain sciences, 40, 2017.
25. Alison Gopnik and Henry M Wellman. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. Psychological bulletin, 138(6):1085, 2012.
26. Jerome H Friedman. The elements of statistical learning: Data mining, inference, and prediction. springer open, 2017.
27. Joaquin Qui˜nonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. Dataset shift in machine learning. Mit Press, 2009.
28. Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pages 80–89.
29. L´eonBottou. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010, pages 177–186. Springer, 2010.
30. Herbert A Simon. Rational decision making in business organizations. The American economic review, 69(4):493–513, 1979.
31. Anastasia Kozyreva and Ralph Hertwig. The interpretation of uncertainty in ecological rationality. Synthese, 198(2):1517–1547, 2021.
32. Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In CHI Conference on Human Factors in Computing Systems, pages 1–18, 2022.

33. Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access, 6:52138–52160, 2018.
34. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267:1–38, 2019.
35. Richard E Nisbett and Timothy D Wilson. Telling more than we can know: verbal reports on mental processes. Psychological review, 84(3):231, 1977.
36. Xiyang Hu, Cynthia Rudin, and Margo Seltzer. Optimal sparse decision trees. Advances in Neural Information Processing Systems (NeurIPS), 2019.
37. Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges.Information Fusion, 2021.
38. EykeH¨ullermeierandWillemWaegeman. Aleatoric and epistemic uncertainty in machine learning:An introduction to concepts and methods. Machine Learning, pages 1–50, 2021.
39. Lyle Brenner, Dale Griffin, and Derek Koehler. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. Organizational Behavior and Human Decision Processes, 97:64–81, 05 2005.
40. Hang Zhang and Laurence Maloney. Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. Frontiers in Neuroscience, 6:1, 2012
41. Daniel Kahneman, Andrew M Rosenfield, Linnea Gandhi, and Tom Blaser. Noise: How to overcome the high, hidden cost of inconsistent decision making. Harvard business review, 94 (10):38–46, 2016.
42. Lorrin M. Koran. The reliability of clinical methods, data and judgments. New England Journal of Medicine, 293(14):695–701, 1975
43. Kenneth B. Little. Confidence and reliability. Educational and Psychological Measurement, 21 (1):95–101, 1961.