The Paradox of Dissonant Predictions: A Central Dilemma of Physician-Algorithm Interaction

Jayson S. Marwaha JAYSONM@MED.UMICH.EDU
University of Michigan, USA

Jeff Choi JC2226@STANFORD.EDU
Stanford University, USA

William Yuan WYUAN95@GMAIL.COM
Flagship Pioneering, USA

Gabriel A. Brat GBRAT@HMS.HARVARD.EDU
Harvard University, USA

Abstract

Clinical algorithms have become highly sophisticated and can outperform physicians in many scenarios. Despite the promise of these tools, uptake and appropriate use is variable. One reason may be because superhuman algorithm performance requires it to come in conflict with a physician's judgment. The paradox is that physicians do not know how to effectively incorporate information that conflicts with their existing beliefs or expectations, even if it may steer them toward the right answer. This confusion around how to confront conflicting algorithmic output is a central obstacle to effective physician-algorithm collaboration. Simply providing accurate recommendations is insufficient; algorithms must effectively change physicians' minds when they are incorrect. This requires rethinking algorithmic design, physician training, and physician-algorithm collaborative models. Rethinking the human-algorithm interface through structured interaction protocols may offer a promising approach. Ultimately, optimizing physician-algorithm synergy likely requires addressing the dissonance generated by a strong model to promote effective integration of algorithmic insights into clinical decision-making.

Keywords: Predictive Processing, Predictive Dissonance, Algorithmic Thinking, Human Computer Interaction

1. Introduction

In 1997 after chess grandmaster Garry Kasparov lost a match to IBM's Deep Blue, he was inspired to study how humans and computers could optimally collaborate to play the game. He began to notice a peculiar trend in his studies: less skilled human players collaborating with a computer often outperformed highly skilled human players collaborating with a computer. The difference between these two groups was that the less skilled players had a well-defined process to interact with the computer recommendations that the more skilled players did not have. As he describes, the friction between skilled players' strong sense of intuition and the computer's recommendations altered the combined gameplay in unpredictable and often harmful ways. Thus was born Kasparov's Law: a weak human working with a machine via a strong process is superior to a strong human working with a machine via a weaker process.(Kasparov, 2018)

Kasparov's Law offers cogent guidance to the burgeoning field of clinical algorithms. Many such algorithms have been developed with the goal of providing the optimal prediction or recommendation for a patient's diagnosis, outcome, or treatment, but for an algorithm to have clinical impact a physician must first decide how to incorporate its output into their decision-making. This final step, the last mile of AI's integration into clinical practice, has the potential to diminish the impact of even the most accurate algorithm. It is generally assumed that generating the best recommendation for a physician will yield the best decision, but humans can respond to algorithmic recommendations in heterogeneous and unpredictable ways. This leads us to the question and the premise of this piece: how should physicians and algorithms interact to improve clinical decision-making?

2. What is Predictive Dissonance?

Mounting evidence shows that giving physicians access to algorithmic tools and AI decision support systems does not reliably improve their decision-making, even in cases where the tool's performance on a diagnostic or prediction task far outperforms the human's. In one study on predicting patient outcomes,(Brennan et al., 2019) an algorithm frequently outperforms surgeons in predicting patients' post-surgical complications. Yet when surgeons were given the algorithm's output and asked to update their prediction, their predictive performance did not significantly improve in most cases. This highlights the paradox of dissonant predictions. The greatest value of clinical algorithms is realized when they change a physician's mind toward the right answer. To do so, they must first deliver information that challenges the physician's pre-existing belief. But as we all know, humans are most resistant to thoughts that are discordant with their own.

An extensive literature exists around the idea of predictive processing in cognitive neuroscience that suggests that we are wired for a positive dopamine response when our internal models match an expected event within the environment.(Elliott Wimmer and Büchel, 2019; Summerfield and de Lange, 2014) Even from a young age, we have a developed expectation-based feedback loop where the mismatch between our internal models and reality lead to a complex rewiring process.(Emberson et al., 2015) Clinical decision support tools that correctly contradict our expectations - which are inherently the most valuable kind - lead to an internal conflict with a negative dopamine response. In such a scenario, it can be easier to rationalize why the algorithm is wrong than to process the complexity of updating one's internal algorithm.

Evidence of this phenomenon can be found in other studies as well; a recent study (Goh et al., 2024) found that that the diagnostic performance of a large language model (LLM) far outperformed human physician diagnostic capabilities, but allowing the human to use the LLM as a decision-making "copilot" did not significantly improve their performance.

It is important to acknowledge that there are many other cognitive biases at play during physician-algorithm interaction that influence the ultimate clinical decision. Automation bias (when a physician allows algorithmic output to override their own intuition) and attribution error (when a physician misinterprets specific algorithmic inputs as causally linked to the patient's outcome) are well described phenomena with a contrary effect.(Gaube et al., 2021) However, the latter effects are usually seen later or are associated with decisions that do not involve a clinician's sense of expertise. Additionally, it is appropriate to provide the caveat that use cases and methods that include immediate feedback loops are much less susceptible to any cognitive biases. Physicians' limited capacity for complex Bayesian reasoning (many studies confirm that physicians perform poorly at estimating probabilities of various outcomes and updating their intuition with new information)(Morgan et al., 2021; Arkes et al., 2022; Manrai et al., 2014) likely also introduces noise into their interactions with algorithms, as they are unsure how to update their clinical judgment appropriately with new algorithmic information.

The underlying theme across all of these phenomena is that incorporating algorithmic predictions is a highly variable process defined by time, expertise, and even personality - especially when it differs from the original expectation. How can we use algorithms in a way that steers our decision-making in the right direction? The path towards a solution for this dilemma likely involves adaptation on the part of both the human and the algorithm.

3. How to Improve Physician-Algorithm Interaction

The goal of physician-algorithm interaction is effective integration of algorithmic recommendations into human judgment: ensuring that the tool steers the user in the right direction when its recommendation is correct, while mitigating any negative effects of incorrect recommendations. There are three possible ways to achieve this. The first involves algorithmic design that, among other things, builds explainability measures or human adjustment levers into models. The second involves rethinking the human-algorithm interface by developing new processes for collaboration. The final approach involves training physicians to be better calibrated to algorithmic recommendations.

3.1. Adapting Algorithms

One proposed strategy for better algorithmic design is to incorporate explainability. Many have argued that designing explainability into an algorithm's output may offer clarity on how the physician should in-

terpret it and therefore most appropriately influence their decision-making. While this is an intuitively appealing concept, current explainability resources do not consistently appear to improve decision-making, possibly because physicians are not trained to fully understand or be convinced by these efforts, and because they can be interpreted in many different ways.(Ghassemi et al., 2021) For example, Jabbour et al.(Jabbour et al., 2023) found that providing heatmaps on chest x-rays that help to explain how the algorithm arrived at its prediction did not mitigate the negative effects of incorrect algorithmic predictions on the radiologists reading those chest x-rays. Until new methods of explainability emerge, it is unlikely to be an effective solution on its own.

Others have suggested designing algorithms that give some amount of control to the physician to adjust the model's output based on their intuition. For example, they may increase the algorithm's predicted risk of a complication if they think they know something about the patient that the algorithm does not capture. While this design approach increases humans' openness to using algorithmic output, there is evidence that it may hurt overall predictive performance.(Dietvorst et al., 2018; Berrigan et al., 2024; Marwaha et al., 2023)

3.2. Novel Collaboration Protocols

Re-imagining the processes by which humans and algorithms interact beyond the traditional unstructured "copilot" model may be a more promising approach. Staying true to the findings of Kasparov, one approach is to establish clear guidelines for physician-algorithm interactions. It has been shown that novice physicians paired with an algorithm via a stronger interaction protocol outperform the diagnostic ability of expert physicians paired with the same algorithm via a weaker interaction protocol; for example, interaction protocols that combined the predictions of "weaker" radiologists and algorithms by first weighting their predictions based on their respective historical accuracy and self-reported confidence levels outperformed "stronger" radiologists who did not interact with an algorithm in this way.(Cabitza et al., 2021)

The ideal collaborative model must also proactively identify scenarios where the human has valuable context about the patient to add to the algorithm. While we previously noted that human adjustment of algorithmic output does not reliably improve decision-making, one longstanding theory for this is because physicians tend to overestimate their understanding of the clinical scenario relative to the algorithms', resulting in excessive adjustment.(Meehl, 1957) To more precisely incorporate valuable human input, we must identify cases in which, first, an algorithm's limited contextual information negatively affects its predictive performance, and second, where the addition of human private data effectively fills that gap. There are scenarios where algorithms can selectively incorporate human input into a prediction when the human knows something the algorithm does not.(Alur et al., 2024)

Another solution is to avoid predictive dissonance altogether by delegating specific tasks to humans or algorithms alone. An algorithm could be allowed to make a decision independently in cases where it is known to perform exceptionally well, and cases with less algorithmic certainty could be delegated to a human expert. One study (Agarwal et al., 2023) showed that the task of interpreting chest x-rays could be safely split between algorithms and radiologists with no degradation in performance and a significant improvement in efficiency. Similarly, algorithms could be allowed to analyze data to suggest a few diagnoses or treatment plans (a task they are known to excel at), while the human clinician could focus on collecting this data from the real world to inform the algorithm via patient conversation and examination and select the most appropriate diagnosis and treatment plan based on this context, tasks that algorithms do not yet perform well.

3.3. Physician Calibration

Algorithmic autonomy is unlikely to be adopted in the real world anytime soon for many good reasons including liability and limitations of algorithmic performance. If ultimately there will always be a "human in the loop", we need approaches for how to make this human better able to identify the right answer. One solution may simply be familiarity and trust of algorithms, earned through repeated empiric validation. As algorithms increasingly become available in clinical settings and physicians see their feedback in the real world, their comfort with incorporating algorithmic output as an element of decision-making will naturally increase. Some have proposed allowing clinicians to use digital tools in risk-free simulated environments prior to deployment to familiarize them-

selves with its strengths and limitations and use it more effectively.(Patil et al., 2025)

More active approaches to expedite familiarity with models could also be taken. More advanced probabilistic thinking can be explicitly taught in medical schools, for example, by incorporating probabilities in case discussions and practicing algorithmic output interpretation.(Goodman et al., 2023) Leveraging existing educational strategies to teach trainees the differences between intuitive versus analytical cognitive processing pathways and how to recognize cognitive biases in their clinical reasoning may also prime them to be more receptive to conflicting pieces of information.(Richards et al., 2020) Classic medical school exposure to case scenarios could include practice using advanced algorithms, so students may become calibrated to the nuances, strengths, and weakness of tools in zero-risk environments.

4. Conclusion

Pablo Picasso once said, "computers are useless - they only give you answers." The last mile of AI's integration into clinical practice will be figuring out how to incorporate these answers productively into our decision-making. As Kasparov observed, collaborative heuristics are critical to human-AI synergy. Conflicts will arise when the algorithm outperforms the clinician. Yet these are the scenarios that justify the use of these tools. Mastering the dance between human and algorithm will require not only retooling well-performing algorithms to be more compelling to humans and creating collaborative protocols that adapt to human biases, but equally importantly we must train physicians to be calibrated to using algorithms by redesigning education around how they engage in clinical reasoning and probabilistic thinking.

5. Citations and Bibliography

References

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. *SSRN Electronic Journal*, July 2023.

Rohan Alur, Loren Laine, Darrick K Li, Dennis Shung, Manish Raghavan, and Devavrat Shah. Integrating expert judgment and algorithmic decision making: An indistinguishability framework. *arXiv [cs.LG]*, October 2024.

Hal R Arkes, Scott K Aberegg, and Kevin A Arpin. Analysis of physicians' probability estimates of a medical outcome based on a sequence of events. *JAMA network open*, 5(6):e2218804, June 2022.

Margaret T Berrigan, Brendin R Beaulieu-Jones, Jayson Marwaha, Anne Fladger, Mark R Katlic, and Gabriel A Brat. Integrating human intuition into prediction algorithms for improved surgical risk stratification. *Annals of surgery*, 279(1):15–16, January 2024.

Meghan Brennan, Sahil Puri, Tezcan Ozrazgat-Baslanti, Zheng Feng, Matthew Ruppert, Haleh Hashemighouchani, Petar Momcilovic, Xiaolin Li, Daisy Zhe Wang, and Azra Bihorac. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. *Surgery*, 165(5):1035–1045, May 2019.

Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. Studying human-AI collaboration protocols: the case of the kasparov's law in radiological double reading. *Health information science and systems*, 9(1):8, December 2021.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3): 1155–1170, March 2018.

G Elliott Wimmer and Christian Büchel. Learning of distant state predictions by the orbitofrontal cortex in humans. *Nature communications*, 10(1):2554, June 2019.

Lauren L Emberson, John E Richards, and Richard N Aslin. Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *Proceedings of the National Academy of Sciences of the United States of America*, 112(31):9585–9590, August 2015.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj digital medicine*, 4(1):31, February 2021.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital health*, 3(11):e745–e750, November 2021.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P J Olson, Adam Rodman, and Jonathan H Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial: A randomized clinical trial. *JAMA network open*, 7(10): e2440969, October 2024.

Katherine E Goodman, Adam M Rodman, and Daniel J Morgan. Preparing physicians for the clinical algorithm era. *The New England journal of medicine*, 389(6):483–487, August 2023.

Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S Valley, Ella A Kazerooni, Nikola Banovic, Jenna Wiens, and Michael W Sjoding. Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. *JAMA: the journal of the American Medical Association*, 330(23):2275–2284, December 2023.

Garry Kasparov. *Deep thinking: Where machine intelligence ends and human creativity begins*. John Murray, London, England, 2018.

Arjun K Manrai, Gaurav Bhatia, Judith Strymish, Isaac S Kohane, and Sachin H Jain. Medicine's uncomfortable relationship with math: calculating positive predictive value: Calculating positive predictive value. *JAMA internal medicine*, 174(6): 991–993, June 2014.

Jayson S Marwaha, Brendin R Beaulieu-Jones, Margaret Berrigan, William Yuan, Stephen R Odom, Charles H Cook, Benjamin B Scott, Alok Gupta, Charles S Parsons, Anupamaa J Seshadri, and Gabriel A Brat. Quantifying the prognostic value of preoperative surgeon intuition: Comparing surgeon intuition and clinical risk prediction as derived from the american college of surgeons NSQIP risk calculator. *Journal of the American College of Surgeons*, 236(6):1093–1103, June 2023.

Paul E Meehl. When shall we use our heads instead of the formula? *Journal of counseling psychology*, 4(4):268–273, 1957.

Daniel J Morgan, Lisa Pineles, Jill Owczarzak, Larry Magder, Laura Scherer, Jessica P Brown, Chris Pfeiffer, Chris Terndrup, Luci Leykum, David Feldstein, Andrew Foy, Deborah Stevens, Christina Koch, Max Masnick, Scott Weisenberg, and Deborah Korenstein. Accuracy of practitioner estimates of probability of diagnosis before and after testing. *JAMA internal medicine*, 181(6):747–755, June 2021.

Shefali V Patil, Christopher G Myers, and Yemeng Lu-Myers. Calibrating AI reliance-a physician's superhuman dilemma. *JAMA health forum*, 6(3): e250106, March 2025.

Jeremy B Richards, Margaret M Hayes, and Richard M Schwartzstein. Teaching clinical reasoning and critical thinking: From cognitive theory to practical application. *Chest*, 158(4):1617–1628, October 2020.

Christopher Summerfield and Floris P de Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature reviews. Neuroscience*, 15(11):745–756, November 2014.

Acknowledgments

Research supported by NIH under award number 3OT2OD032581-01.