MACHINE ASSISTED HUMAN DECISION MAKING

Sara Mourad and Ahmed Tewfik

Department of Electrical and Computer Engineering
The University of Texas at Austin

***Abstract*—Artificial intelligence is often touted as the ultimate automation technology capable of outperforming humans. It is also feared by some because of its potential to eliminate certain jobs. In this paper, we describe scenarios in which man-machine symbiosis, or properly designed combinations of man and machine, can actually outperform man and machine. We also present a statistical solution to a constrained version of a generic problem in man-machine symbiosis. Specifically, we solve the problem of optimal selection, ordering and presentation of data to a human to solve a class of problems that artificial intelligence can fail to solve on its own, such as fraud detection. The man-machine symbiosis solution we present overcomes human cognitive biases which stand in the way of their rational decision making.**

Index — Cognitive biases, Subset selection, Bayesian Hypothesis Testing, Man-Machine.

I. INTRODUCTION

Man-machine symbiosis can outperform man and machine alone in several tasks, like fraud detection. This is because humans have some expertise and experience that cannot be fully learned by the machine. On the other hand, humans are not rational decision makers. They are subject to cognitive biases; heuristics that can lead them to wrong conclusions [12]. For instance, large scale problems cannot be handled by humans optimally. We propose in this paper man-machine symbiosis where the human experts are the decision makers and the machine mitigates cognitive biases to transcend human limitations.

Several cognitive biases have been studied in the literature [6]. As an example of cognitive biases, we cite the anchoring bias where humans are influenced by starting points or initial beliefs. In [7] and [13] the starting point bias is modeled by the impact of the initial bid value on the willingness to pay. Other biases are the confirmation bias where humans tend to emphasize observations confirming their belief, and neglect observations contradicting their beliefs. In [4], the belief update model is modified to account for the confirmation bias in the context of auditing. The overconfidence bias is investigated in [10], where they verify its existence in analyst earning forecasts. In [14], Hogarth and Einhorn study the order effects in the update of belief. More precisely, they point out that humans can be subject to the primacy effect [3, 22] where humans emphasize the first set of information, and the recency effect where the last set of information is emphasized. They show also that with an increasing amount of data presented to them, subjects can get tired and become less sensitive to new observations [14]. Moreover, in [25], they show experimentally that different order in which information is presented to humans can lead to opposite decisions.

As part of the efforts to do cognitive bias mitigation, [9] proposes a serious video game that detects some biases in the players' performance and teaches them how to identify and mitigate them through the use of feedback. [5, 8] also uses serious games as a way to recognize and mitigate human biases including the anchoring bias. An alternative to training humans in order to reduce biases in their decisions is to modify the way decisions are presented to them. In experiments that involved participants to play a game or watch an instructional video, [19] proposes giving a single training intervention during the experiments as an effective way to mitigate the biases addressed in these experiments. In [21], the Bayesian updating model is modified to incorporate and model the cognitive biases in human decision making. Based on this model for human information processing, [20] studies the problem of optimally ordering data to human subjects in order to mitigate biases using the Neyman-Pearson test [11] as the statistical test for decision making. In this paper, we use the Bayesian hypothesis test [17], and we study the same problem of optimal data ordering, but we allow the incorporation of human expertise that is unknown to the process ordering the data, as well as the prior probabilities of the possible hypotheses. In [2], the human bias is considered under the sequential probability ratio test (SPRT) [17]. In this different setting, data is shown to the human subject until the latter reaches a confident decision, and discards the subsequent information. In other words, the problem considered wouldn't include any selection of the data, but rather the optimal ordering of data to mitigate biases under this different setting. In [1], the bias is modeled by modifying the thresholds in the SPRT, and ordering is done based on the statistics of the data, and not on the sufficient statistics as done in this paper.

In this paper, we investigate the problem of optimally combining the man and machine, by studying how a machine can optimally select and order observations to humans in polynomial time, under the general framework where humans know additional information through their expertise that the machine has no access to. In Sec. II of the paper, we describe scenarios in which human machine collaboration can outperform human and machine alone. In Sec. III, we review the model used for human information processing, and present the problem of optimal ordering of data to show to humans, as well as its solution. In Sec. IV, we show through simulations

that the proposed approach allows human performance to be near optimal under the confirmation bias, and that the approach is robust to outliers. And finally in Sec. V, we review the challenges of machine assisted human decision making.

II. THE IMPORTANCE OF COLLABORATION BETWEEN MAN AND MACHINE IN DECISION MAKING

Moravec's paradox states that humans and machines are good at different tasks [23]; men are better at intuitive tasks while machines are better in more resource intensive tasks. In 2005, the online chess-playing site Playchess.com hosted a free style tournament where teams could use the help of computers [15]. The surprising outcome was not that the winner team included both human players and computers, but that a pair of amateur players using three computers were able to outperform a grandmaster with the most updated computer. This was possible because of the way they coached their computers to look very deeply into positions [15]. Therefore the takeaway is not only that men with machine can outperform either alone, but that we should also figure out what is best way they can complement each other.

Fraud detection is another example where human intervention is needed. Specifically, machine learning can only capture fraud that is similar to what it has seen or been trained on in the past. But attackers are creative and always try new tactics. While machine learning problems will learn eventually the new trends, they would need more substantial data from the new attack in order to learn it. Machine learning algorithms would be ideal to process the large amount of data, and to detect the obvious cases. But uncertain cases should be reviewed by experts who can use their skills to make the right decision, or further expand the investigation and ask for more data for a specific case. And so to go beyond the overall detection rate of machines, one needs to determine what to show to a human being and in what order to maximize the chance that the human being will pick up the new fraud mechanism. Let's consider for example that the human subject is asked to review a decision of "no fraud" by the machine. First, there is the risk of anchoring biasing the human with the initial judgment of "no fraud". Also, the amount of features to show to the human subject could be very large. Hence, we need to decide which observations, and in which order, should be shown to the human subject so that we mitigate their potential biases.

III. COGNITIVE BIASES IN DECISION MAKING

In order to illustrate man-machine symbiosis, we consider decision-making in a simple binary hypothesis testing problem. The binary hypothesis testing problem decides between two hypotheses H_0 and H_1 based on the vector Y of N observations y_i, $0 \leq i < N$. In the continuous-valued case, $Y \in R^N$, and Y admits the following probability densities:

$$\begin{aligned} H_0 &: Y \sim f(Y|H_0) \\ H_1 &: Y \sim f(Y|H_1), \end{aligned} \tag{1}$$

Assuming the observations are independent, let $f(.|H_i)$ denote the probability density function under hypothesis H_i, and let l_i denote the log-likelihood ratio for observation y_i where $l_i = \log(\frac{f(y_i|H_1)}{f(y_i|H_0)})$, and L_i the cumulative log-likelihood ratio up to the ith observation y_i, such that $L_i = \log(\prod_{k=1}^{i} \frac{f(y_k|H_1)}{f(y_k|H_0)}) = \sum_{k=1}^{i} l_k$.

A. Biased information processing model

As suggested in [14], and in order to model the cognitive biases of human beings, [21] introduces a modification of the traditional Bayesian updating by introducing adjustment weights w_i when calculating the cumulative log likelihood ratio. The Bayesian updating model is modified as follows:

$$L_i^b = L_{i-1}^b + w_i l_i \tag{2}$$

where L_i^b denotes the biased cumulative log likelihood ratio, w_i is the adjustment weight that the subject gives to the new observation due to biases (w_i depends on l_i or L_i^b). For example, the confirmation bias is modeled by giving high adjustment weights to data confirming the hypothesis to which a human subject is biased, and low adjustment weights to disconfirming data. This model is based on the anchoring and adjustment model proposed and validated by Hogarth and Einhorn[14], which states that humans update their beliefs by subjectively weighing a new observation depending on their current belief at the time of its acquisition.

B. Problem statement

In the context of binary hypothesis decision-making, consider a human Bob subject to cognitive biases. Consider also a vector Y of N independent observations, and a machine used to select and order N' out of the N observations to show to Bob. The machine has only access to Y. Bob, and aside of the N' observations which will be presented to him, knows a set of observations X through his past experience and other past events he is aware of. Since Bob will be using a Bayesian hypothesis test to decide the hypothesis, he will be setting a specific threshold λ based on his estimates of the costs and the prior probabilities. The problem is to find out which N' observations, and in which order, should the machine present to Bob such that his performance is optimal. We define an optimal performance as being that of an oracle who knows X (known to Bob), the N-sized vector Y (known to the machine), and the threshold λ (set by Bob). We define $L_B = L_X + L_{N'}^b = L_X + \sum_{i=1}^{N'} w_i l_i$, as the cumulative log likelihood ratio of Bob, where L_X is the cumulative log likelihood ratio of Bob based on

observations X before seeing any observation of Y, and $L^b_{N'}$ is the biased cumulative log likelihood ratio based on the N' sized ordered subset of the observations Y. We also define $L_O = L_X + L_N = L_X + \sum_{i=1}^{N} l_i$, as the cumulative log likelihood ratio of the oracle, where L_N is the cumulative log likelihood ratio based on the observations Y. We note that L_X here is assumed to remain constant as the human receives the observations y_i. Dealing with L_X changing with the observations y_i is out of the scope of the paper, and would require interaction with the human being to estimate it. We also note that a similar problem has been considered in [20], with the distinction that this paper, unlike [20], is able to capture the man-machine symbiosis by incorporating the past knowledge and the expertise of the human subject. This has been possible by using the Bayesian hypothesis framework instead of the Neyman-Pearson test. Moreover, this framework allows to account for the priors of the hypotheses (or possibly biased priors), as well as to generalize the approach for any distribution of the independent observations Y.

C. Optimizing human decision making

The Bayesian hypothesis test solves the binary hypothesis testing problem by choosing the hypothesis which minimizes the average cost of the decision, also known as the Bayes risk: $R = \sum_{i=0}^{1}\sum_{j=0}^{1} C_{ij} P(H_i|H_j) P_j$, where C_{ij}, $0 \leq i, j \leq 1$, represents the cost of deciding H_i when H_j is the true hypothesis, and the prior probability of hypothesis H_0 is P_0 and of hypothesis H_1 is P_1. The optimal test becomes as follows:

$$L \underset{<}{\geq} \lambda = \log(\tau \frac{(C_{10} - C_{00})}{(C_{01} - C_{11})}) \tag{3}$$

where L denotes the cumulative log likelihood ratio at the time of decision, and $\tau = \frac{P_0}{P_1}$. Since we need to match the performance of Bob to that of the oracle, the problem boils down to:

$$\underset{\mathcal{K}\subseteq[N]:|\mathcal{K}|=N'}{\operatorname{argmin}} |R(L_O) - R(L_B)| \tag{4}$$

where

$$\begin{aligned} R(L_O) = {} & C_{10}P_0 P(L_O > \lambda|H_0) + C_{01}P_1 P(L_O < \lambda|H_1) \\ & + C_{00}P_0 P(L_O < \lambda|H_0) + C_{11}P_1 P(L_O > \lambda|H_1) \end{aligned} \tag{5}$$

is the Bayes risk of the oracle and

$$\begin{aligned} R(L_B) = {} & C_{10}P_0 \hat{P}(L_B > \lambda|H_0) + C_{01}P_1 \hat{P}(L_B < \lambda|H_1) \\ & + C_{00}P_0 \hat{P}(L_B < \lambda|H_0) + C_{11}P_1 \hat{P}(L_B > \lambda|H_1) \end{aligned} \tag{6}$$

is the Bayes risk of Bob, and where $\hat{P}$ denotes the altered statistics of the cumulative log likelihood ratio of the ordered subset $\mathcal{K}$ of N' observations after selecting them from the complete set of N observations.

Let $Pd_O(\lambda)$ ($Pd_B(\lambda)$) and $Pf_O(\lambda)$ ($Pf_B(\lambda)$) denote the probabilities of detection and false alarm of the oracle (Bob), respectively. The problem can be rewritten as :

$$\begin{aligned} \underset{\mathcal{K}\subseteq[N]:|\mathcal{K}|=N'}{\operatorname{argmin}} & |(C_{10}P_0 - C_{00}P_0)(Pf_O(\lambda) - Pf_B(\lambda)) \\ & - (C_{01}P_1 - C_{11}P_1)(Pd_O(\lambda) - Pd_B(\lambda))| \end{aligned} \tag{7}$$

This problem is solved when matching the tail probabilities of Bob and the oracle without needing to figure out $\hat{P}$, but instead by making L_B as close as possible to L_O. And now if the ordered subset $\mathcal{K}$ of N' elements is chosen such that

$$L_X + L^b_{N'} - \lambda = L_X + L_N - \lambda \tag{8}$$

and equivalently $L^b_{N'} = L_N$, then the tail probabilities are matched. Therefore, we are interested in finding an ordered subset $\mathcal{K}$ of N' observations such that $L^b_{N'}$ is as close as possible to the target $T = L_N$. In fact, using techniques from [20], we can show that if $L^b_{N'}$ is such that $T - \frac{\delta}{2} \leq L^b_{N'} \leq T + \frac{\delta}{2}$, then $Q(\frac{\lambda - L_X - \mathbb{E}[L_N|H_1] + \frac{\delta}{2}}{\mathrm{Var}(L_N)}) \leq Pd_B(\lambda)$ and $Pf_B(\lambda) \leq Q(\frac{\lambda - L_X - \mathbb{E}[L_N|H_0] - \frac{\delta}{2}}{\mathrm{Var}(L_N)})$ for Gaussian observations.

We note that here, we assumed that the priors of Bob are the true ones. If Bob has biased priors (i.e. the cognitive biases are not only incurred by the presented observations, but are present even before), then $\tau_b = \eta \frac{P_0}{P_1}$, where τ_b is the biased ratio of Bob's priors, and η is a multiplicative factor that incorporates the bias. And so if we define $\hat{L}^b_{N'} = L^b_{N'} - \log \eta$, the problem can be solved similarly by replacing $L^b_{N'}$ with $\hat{L}^b_{N'}$, and hence we would need $\hat{L}^b_{N'} = L_N$, i.e. $L^b_{N'} = L_N + \log \eta = T$.

IV. RESULTS

The problem boils down to finding in polynomial time an ordered subset whose weighted sum is within a guaranteed small distance of a given target. This setup is the same as the one in [20], and hence we use their proposed approximation algorithm in order to find the subset, which is a modification of the approximate subset sum algorithm [18]. The solution returned is within a factor $1 + \epsilon$ of the optimal solution. The running time is polynomial in both N and $\frac{1}{\epsilon}$.

A. Confirmation bias and results

We test our algorithm under the setting described in Eq. 1 and using Gaussian distributions. In modeling the confirmation bias, the adjustment weight w_i in Eq. 2 depends on the value of the current log likelihood ratio l_i. We model the bias by assigning small adjustment weights w_i to log likelihood ratios l_i contradicting a given hypothesis, and w_i close to 1 for l_i supporting this hypothesis.

We simulate the performance of the algorithm whenever the hypothesis supported by the subject is H_0. We repeat the simulations for different ratios $\frac{N'}{N}$, i.e. for different

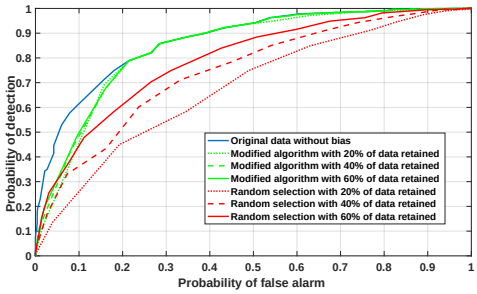

Figure 1: ROC curves under different selection of Gaussian observations with confirmation bias

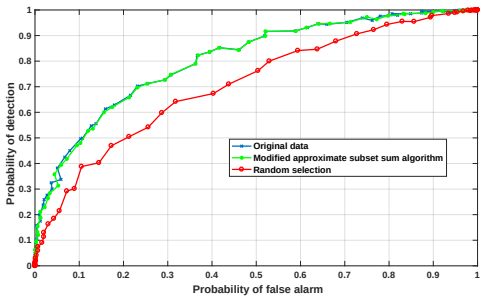

Figure 2: ROC curves under different selection of Gaussian independent observations with presence of outliers

percentages of the data selected. In the simulations, the percentages of data retained take the values 20%, 40% and 60%. As shown in Fig. 1, and for all values of the ratio $\frac{N'}{N}$, the algorithm gives ROC curves (green) nearly overlapping the optimal ROC curve (blue) for relatively high probability of detection. However, for low probability of detection, the green ROC curves don't completely match the original blue curve. This is because whenever hypothesis H_1 is true, the bias (towards H_0) leads to larger gap between the closest subset sum to the target and the target, and when the probability of detection is low, it is more important for the subset sum to be as close as possible to the target. Also, as shown in Fig. 1, the higher the ratio $\frac{N'}{N}$, the closer the green ROC curves to the original blue ROC curve (only noticeable for low probability of detection), and the closer the red ROC curves of the random selection to the original ROC. In the random selection case, the higher the ratio $\frac{N'}{N}$, the closer the statistics of $L^b_{N'}$ to the statistics of the original L_N, and therefore the closer the performance of the random selection to the oracle's performance. Now when the selection algorithm is applied, it is beneficial to increase the ratio $\frac{N'}{N}$ so that there is a higher number of possible combinations resulting in $L^b_{N'}$, and thus a higher probability for $L^b_{N'}$ to be close to L_N.

B. Outliers

It is important to check that the presence of outliers in the original data won't affect the performance of the selection algorithm; the outliers may have little effect on the large number of observations, and we want to make sure that this still holds after cutting down on the number of observations. We expect the proposed algorithm to be robust to outliers even with using less observations, because the algorithm matches $L^b_{N'}$ to L_N, and thus the advantage of having all the data (and therefore knowing L_N) is preserved after selecting the observations. To verify that, we run simulations where with a small probability, we generate an outlier observation from a Gaussian distribution with a mean far from the mean of the Gaussian distribution of the corresponding hypothesis. $\frac{N'}{N}$ is set to be 40%. As shown in Fig. 2, the ROC obtained using the proposed algorithm still matches the performance of the optimal ROC using the whole set of observations N.

V. THE CHALLENGES OF HUMAN MACHINE INTERACTION: ASYMMETRIC INFORMATION

A. Learning humans

Human machine collaboration presents many challenges and open problems. The challenges mainly lie in learning the human behavior. In Eq. 8 , L_X could change with each new observation presented to the human subject. Cases when L_X could vary with new observations are for example when humans recall additional past observations when shown a certain observation. However, the mechanism in humans that lead them to remember information based on the presented information to them is very complex. It is based on the relation between their past experiences and the new pieces of information shown to them. Hence ordering data to trigger the recall of information in the human brain that is pertinent to the problem but unknown to the machine is a difficult problem to which we don't know the solution yet.

Another unaddressed problem is that the weights parameters in our model in Eq.2 should be learned through an adaptive process that would require interaction with the human subjects by asking them to do specific tasks.

B. Black box models

On the other hand, in the context where the human receives the decision from a machine and has to make the final decision, the human often doesn't understand why the decision has been made. With state of the art models like neural networks, these algorithms are black boxes to humans beings, which limits the value of the help that the human is getting from the machine. A lot of effort is being spent in understanding which features lead to this decision, and the reason behind those decisions, leading to more interpretable machine learning models [16, 24].

REFERENCES

[1] N. Akl and A. Tewfik. Optimal information ordering in sequential detection problems with cognitive biases. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1876–1880, May 2014.

[2] N. Akl and A. Tewfik. Optimal information ordering for sequential detection with cognitive biases. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 413–417. IEEE, 2016.

[3] E. M. Altmann. Memory in chains: Modeling primacy and recency effects in memory for order, 2000.

[4] E. Bamber, R. J. Ramsay, and R. M. Tubbs. An examination of the descriptive validity of the belief-adjustment model and alternative attitudes to evidence in auditing. *Accounting, Organizations and Society*, 22(34):249 – 268, 1997.

[5] M. Barton, C. Symborski, M. Quinn, C. K. Morewedge, K. S. Kassam, and J. H. Korris. The use of theory in designing a serious game for the reduction of cognitive biases. *Transactions of the Digital Games Research Association*, 2(3), 2016.

[6] M. H. Birnbaum and B. A. Mellers. Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, 45(4):792, 1983.

[7] Y.-L. Chien, C. J. Huang, and D. Shaw. A general model of starting point bias in double-bounded dichotomous contingent valuation surveys. *Journal of Environmental Economics and Management*, 50(2):362–377, 2005.

[8] B. A. Clegg, B. McKernan, R. M. Martey, S. M. Taylor, J. Stromer-Galley, K. Kenski, E. T. Saulnier, M. G. Rhodes, J. E. Folkestad, E. McLaren, et al. Effective mitigation of anchoring bias, projection bias, and representativeness bias from serious game-based training. *Procedia Manufacturing*, 3:1558–1565, 2015.

[9] N. E. Dunbar, M. L. Jensen, C. H. Miller, E. Bessarabova, S. K. Straub, S. N. Wilson, J. Elizondo, J. K. Burgoon, J. S. Valacich, B. Adame, et al. Mitigating cognitive bias through the use of serious games: Effects of feedback. In *International Conference on Persuasive Technology*, pages 92–105. Springer, 2014.

[10] G. Friesen and P. A. Weller. Quantifying cognitive biases in analyst earnings forecasts. *Journal of Financial Markets*, 9(4):333–365, 2006.

[11] R. Gallager. Detection, decisions, and hypothesis testing, 2012.

[12] T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

[13] J. A. Herriges and J. F. Shogren. Starting point bias in dichotomous choice valuation with follow-up questioning. *Journal of Environmental Economics and Management*, 30(1):112 – 131, 1996.

[14] R. M. Hogarth and H. J. Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1 – 55, 1992.

[15] G. Kasparov. The chess master and the computer. *The New York Review of Books*, 57(2):16–19, 2010.

[16] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

[17] B. C. Levy. *Principles of Signal Detection and Parameter Estimation*. Springer US, 2008.

[18] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

[19] C. K. Morewedge, H. Yoon, I. Scopelliti, C. W. Symborski, J. H. Korris, and K. S. Kassam. Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):129–140, 2015.

[20] S. Mourad and A. Tewfik. Real-time data selection and ordering for cognitive bias mitigation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4393–4397, March 2016.

[21] S. Mourad and A. H. Tewfik. Cognitive biases in bayesian updating and optimal information sequencing. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 4095–4099. IEEE, 2015.

[22] R. Nisbett and L. Ross. *Human inference: strategies and shortcomings of social judgment*. Century psychology series. Prentice-Hall, 1980.

[23] D. Rasskin-Gutman. *Chess metaphors: artificial intelligence and the human mind*. MIT Press, 2009.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[25] X. Zhu, B. R. Gibson, K.-S. Jun, T. T. Rogers, J. Harrison, and C. Kalish. Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1247–1254, 2010.