Enhancing Fairness, Justice and Accuracy of Hybrid Human-AI Decisions by Shifting Epistemological Stances.

Peter Daish[0009−0006−6812−1791], Matt Roach[0000−0002−1486−5537], and Alan Dix[0000−0002−5242−7693]

Swansea University, Swansea SA1 8EN, UK
peter.daish@swansea.ac.uk*
m.j.roach@swansea.ac.uk
a.j.dix@swansea.ac.uk

Abstract. From applications in automating credit to aiding judges in presiding over cases of recidivism, deep-learning powered AI systems are becoming embedded in high-stakes decision-making processes as either primary decision-makers or supportive assistants to humans in a hybrid decision-making context, with the aim of improving the quality of decisions. However, the criteria currently used to assess a system's ability to improve hybrid decisions is driven by a utilitarian desire to optimise accuracy through a phenomenon known as 'complementary performance'. This desire puts the design of hybrid decision-making at odds with critical subjective concepts that affect the perception and acceptance of decisions, such as fairness. Fairness as a subjective notion often has a competitive relationship with accuracy and as such, driving complementary behaviour with a utilitarian belief risks driving unfairness in decisions. It is our position that shifting epistemological stances taken in the research and design of human-AI environments is necessary to incorporate the relationship between fairness and accuracy into the notion of 'complementary behaviour', in order to observe 'enhanced' hybrid human-AI decisions.

Keywords: Epistemologically Driven Hybrid Human-AI Environment Design · Human-AI interaction · Human-AI Fairness, Justice and Accuracy.

1 Introduction

Typical approaches to hybrid human-machine learning or human-AI (Artificial Intelligence) decision-making systems are often developed with the intent of observing accuracy enhancing 'complementary behaviour' [9]. Whilst the need for accuracy is important for developing trust in hybrid human-AI (and human-Robot) interactions [24, 1], accuracy alone is a misleading assessment criteria for validating decisions, since accurate and inaccurate systems alike can perform undesirably for underrepresented groups [12].

Indeed, AI4People's "Ethical Framework for a Good AI Society" [4] recommended to the EU parliament that AI is validated according to its satisfaction of key factors, such as: Non-maleficence, Beneficence, Autonomy, Justice and Explicability rather than just Accuracy. However, the authors made a commonly held, but controversial, assumption that fairness and justice are interchangeable terms: Scholars in organisational behaviour present historical evidence to show that that fairness and justice are fundamentally different [7]. Justice refers to rule adherence whilst fairness refers to an individual's response to the moral perception of those rules [7]. Furthermore, the relationship between fairness and justice is neither mutually necessary, nor mutually exclusive, for example: a judicial decision incurring a penalty can be considered just according to the law and a fair penalty by the judge given the actions of the defendant, but unfair according to the defendant's moral perception of the law, how and why it was broken and the extent of the incurred penalty. As such, we stipulate that fairness ought to be included as a fifth pillar in the AI4People's framework for validating human-AI systems, rather than assumed to be satisfied under the condition of Justice. We believe that making the distinction between fairness and justice clear will steer researchers towards a better understanding of how hybrid human-AI decision making systems are perceived to perform in a just and fair manner, with differing levels of automation and decision support.

It is also our belief that the positioning of the AI within the hybrid human-AI decision-making environment potentially limits the resultant fairness of combined decisions. This is the case because AI currently optimise for objective definitions of subjective concepts such as fairness, which are static and limited representations of complex social phenomena. In this position paper we split human-AI hybrid decision-making systems into three categories, denoted by the epistemological stance taken in their design and according to the positioning of the AI within the human-AI environment. By applying the definitions of epistemologies found in [15] we denote hybrid human-AI decision making systems as belonging to one of three categories: objectivist-inspired, subjectivist-inspired and constructivist-inspired.

2 Discussion

2.1 Typical AI powered decision-making limited by an objectivist-inspired epistemological stance.

The typical approach to hybrid human-AI and automated decision-making was guided by an objectivist epistemology and can be found in some of the first breakthroughs in deployed AI technologies - expert systems (ES) - but has also been carried through as the predominant design philosophy for modern data-driven human-AI environments. An example of this environment design can be seen in Figure 1, where a human decision-maker is aided by AI-assisted decision-support through recommendations.

Expert systems acted according to a knowledge base designed to encode decision-logic, as defined by experts in the decision-domain [11, 21, 18]. This ap-

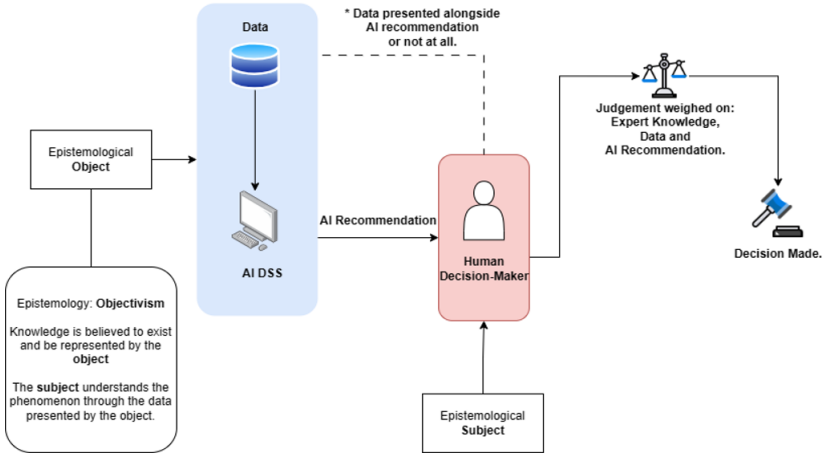

Fig. 1. This figure demonstrates an objectivist epistemological approach to human-AI environment design, which is typical of the field.

proach is inspired by the epistemological stance of objectivism, which stipulates that knowledge (the decision) about an object (the data) is independent of the subject (the decision-maker) [15]. A pragmatic application of this can be found in Turban and Watkins [11], where they refer to ES being developed for "well-structed" problems . Whilst ES could perform well and were found to improve the accuracy of decisions in some contexts [11], their performances were contingent on a manual process of knowledge engineering which depended heavily on interviewing domain-experts in an attempt to elicit the decision-logic [21]. In this example, the positioning of the AI is as an 'artificial expert' within a human-AI decision-making environment and the positioning of the human is one of receiving automated decision-recommendations from an approximate model of encoded domain intelligence. The human decision-maker in this hybrid human-AI environment must manually identify and remediate cases where the AI is found to perform unexpectedly [11].

Concerns surrounding the understandability and interpretability of decisions made through or with ES led to a call-to-action for so-called "second generation" ES to incorporate human factors into ES design. It was hoped that, by re-introducing subjectivist and constructivist inspired concepts, such as explainability, interaction and co-operation, we could observe complementary performance and greater trust in hybrid human-AI decisions. However, these systems, due to the objectivist-inspired epistemological stance used in their design to capture and define "knowledge" are susceptible to changing environments, where expert understanding is influenced by social biases and temporal factors, such as dynamic social norms. Indeed, though not explicitly referred to as an expert-system - but rather the colloquial term 'algorithm' - systems using expert domain knowledge encoded in their design have been found to perform in an unjust and unfair manner according to racial characteristics [3]. Whilst human decision-makers may also exhibit bias in their decision-making, these systems

are capable of perpetuating social injustice on a mass-scale through automated and semi-automated decision-making.

The same objectivist epistemologically inspired approach can be observed in many modern hybrid human-AI decision making systems which use data-driven deep-learning (DL) technologies to learn complex relationships between input data and output label by optimising for predictive accuracy. Whilst many of these systems are accurate, their mechanisms are often opaque (blackbox), meaning that explanations as to why the AI made its recommendations are often non-trivial to produce [19, 22]. Additionally, DL technologies are also vulnerable to learning biased representations of phenomena during training, which can lead them to perpetuate systemic biases in their recommendations to human decision-makers [13]. Due to their blackbox nature, these biased recommendations are difficult to assess and account for when used in hybrid decision-making contexts. A growing field of study is attempting to incorporate subjective philosophical notions of "Fairness" into statistically testable definitions and optimisation targets for DL technologies to make their decision-making fairer. Example fairness definitions include adaptations of: demographic parity, equality of odds and equality of opportunity [8]. These papers refer to fairness as mathematically define-able concepts which can be optimised for, or else used as an assessment criteria by DL technologies, but in reality, they are referring to simplified notions of "justice" [7], not fairness.

Defining notions of fairness objectively, is problematic: human perceptions of fairness are dynamic and change with life-experience or with new information [10]. Thus, the notion of fairness as it exists within a decision-maker is not bound to an abstract mathematical formula and can evolve through their life through complex socio-techincal interactions well beyond the scope of fairness enhanced AI-assisted decision support. Moreover, whilst the subjectivity of fairness does not mean that it cannot be defined in objective terms by a subject at a point in time, it is known that objectively defined notions of fairness are incompatible with one another [6, 20]. This means that, even though formulaic notions of fairness can theoretically be derived, for any given decision-making environment wherein multiple users or stakeholders are involved each with their own notions of fairness, there are likely to be conflicts between what's deemed fair and unfair. This means that ultimately, a subjective decision to choose one fairness definition over the other must be made - or else a compromised definition for fairness found across stakeholders - to incorporate fairness in a human-AI environment. In addition, since fairness perceptions can change over time, any decision made over the chosen fairness definition for a human-AI environment would have to be continually updated to reflect the values of its stakeholders.

Additionally, it has been reported that fairness as an objectively derived notion has a competitive relationship with accuracy [14]. Currently, one of the major design goals of human-AI environments is the desire to observe complementary behaviour or 'enhanced' human-AI decisions, however, this notion typically refers to the utilitarian desire to increase human-AI accuracy. The desire for increased 'utility' through human-AI cooperation is thus a conscious

decision which could drive unfairness. We propose re-framing the notion of 'complementary behaviour' to incorporate subjective yet influential factors that affect decision-making and its perception, such as Fairness, in the design of human-AI environments.

Whilst applying objective definitions of fairness in human-AI environments can be problematic and we question the continued method of objectively deriving fairness metrics as optimisation targets for AI or human-AI systems, fairness, justice, accuracy and the position of the human and AI in the hybrid environment have nevertheless been considered as important factors in the acceptance and trust in hybrid human-AI decision-making [17, 23, 5, 2]. As a result, alternative methods for incorporating subjective notions into the human-AI environment ought to be explored. To this extent, we propose shifting design philosophies for human-AI environments such that the environments themselves are designed to cater for the subjective elements of decision-making.

2.2 Towards subjective and constructivist inspired hybrid human-AI designs for Fairness, Justice and Accuracy

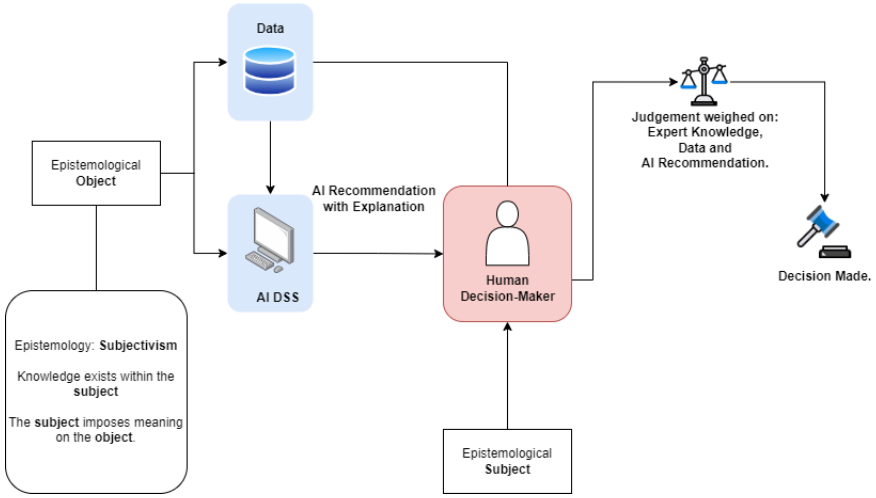

Fig. 2. This figure demonstrates a potential human-led, subjectivist-inspired environment for human-AI decision-making

Adapting to [15]'s definition of subjectivism, a human-led subjectivist-inspired hybrid environment is one where the human is depicted as possessing expertise or knowledge about the underlying data of the decision to be made and the AI is present as a means of supplying useful contextual information relating to the decision. This might be, for example, decision-recommendations alongside explanation techniques such as AI model confidence scores [25] or through explanation of historic precedent - as with Shanghai's 206 system [23]. An example of this

environment can be found in Figure 2. Within this environment, the human decision maker will use their experience and expertise in the decision-domain to formulate a decision, whilst using evidence from explained AI recommendations as supporting evidence to increase their confidence in the decisions. Fairness-explained AI decision-support - which details potential injustices caused by its recommendations - can be used by experts to mitigate biases being perpetuated through AI-assistance in hybrid environments [16]. Alternatively, AI decision-support might track a human's bias profile and alert them to unjust trends in their own decision making behaviour, so that the human decision-makers can self regulate bias in future decisions. In both of these cases algorithmic fairness metrics (which are truly denoting algorithmic justice) are being used as means of identifying injustices within the hybrid environment, as opposed to optimisation targets. This gives the responsibility back to the human-decision maker to reconcile these decisions and recommendations with notions of fairness, given: their interpretation of the assessment criteria, the data provided to them, their prior experience in the domain and also the presence of the AI's support.

Whilst a human-led subjectivist-inspired approach is intuitive to imagine given the development of Explainable AI technologies, it is less clear how an AI-led subjectivist-inspired environment would be designed. However, a possible AI-led environment might include the AI as primary decision-maker, with the human providing input regarding only the subjective aspects of decision-making, such as ranking the fairness of decision-outcomes for a decision to be made.

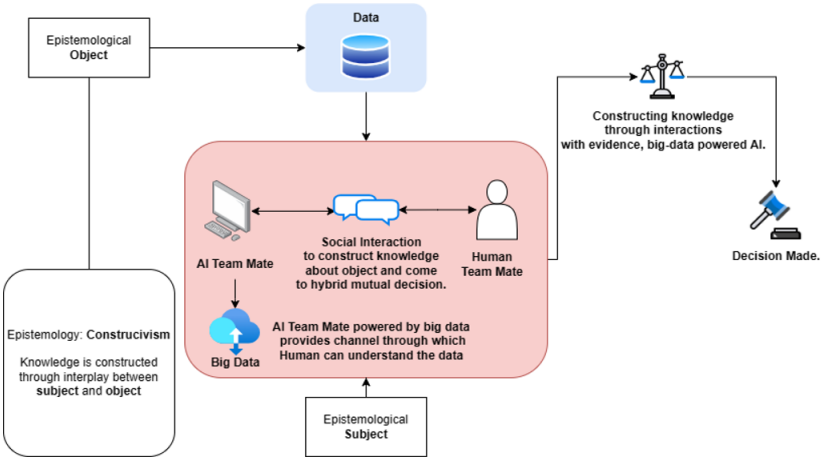

Fig. 3. This figure demonstrates an example constructivist human-AI environment where human and AI are social actors who construct decisions through social interactions with one another.

An alternative approach to the subjectivist-inspired paradigm for future study, would be to use a constructivist-inspired hybrid environment. Adapting from [15], this type of environment might position the human as the decision-

making subject using AI support to interact with the objective data used to make a decision in order to better understand it. An example environment setup for this approach can be seen in Figure 3. In this hybrid environment, the AI acts both as the medium through which the human understands the data and a social collaborator. In the scenario where the human is an expert in the decision-domain, the AI can expose insights into its own recommendation logic through explanation and allow the expert to explore these trends in an interactive manner. In the alternative scenario where the human is a non-expert in the decision domain, the AI would attempt to educate the decision-maker as to why it made its recommendation, providing supporting evidence drawn from precedent. This is similar to Swartout and Moore's [21] second generation systems and the notion of generating personalised recommendations.

2.3 Conclusion and Call To Action

The two alternative epistemological approaches contributed by this paper aim to re-allocate the task of upholding subjective notions of fairness and justice within the decision-making process to the human decision-maker, rather than optimising for them algorithmically. We propose that re-allocating subjective decision-making tasks to humans who are members of society and thus, have their own opinions on the nature of fairness and justice, whilst supporting them by providing accurate recommendations and information that could help apply their perceptions of fairness and justice to the underlying decision data, would play off the strengths of both humans and AI within human-AI environments. We propose a call-to-action for hybrid decision making systems research to explore how alternative epistemologically inspired approaches to their design can enhance the fairness, justice and accuracy of decisions.

Acknowledgements. This work was funded by Swansea University and the Economic and Social Research Council (ESRC) under the grant code ES/P00069X/1. For the purpose of Open Access, we have applied a CC BY license to any Author Accepted Manuscript (AAM) version arising from this submission. Additionally, we would like to thank Keneni Tesema for her patience in listening to and conversing on the principles of this research.

References

1. Alzahrani, A., Robinson, S., Ahmad, M.: Exploring factors affecting user trust across different human-robot interaction settings and cultures. In: Proceedings of the 10th International Conference on Human-Agent Interaction. p. 123–131. HAI '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3527188.3561920, https://doi.org/10.1145/3527188.3561920
2. Bankins, S., Formosa, P., Griep, Y., Richards, D.: AI Decision Making with Dignity? Contrasting Workers' Justice Perceptions of Human and AI Decision Making in a Human Resource Management Context. Information Systems

Frontiers **24**(3), 857–875 (Jun 2022). https://doi.org/10.1007/s10796-021-10223-8, https://doi.org/10.1007/s10796-021-10223-8
3. Braun, L., Grisson, R.: Race, Lung Function, and the Historical Context of Prediction Equations. JAMA Network Open **6**(6), e2316128–e2316128 (06 2023). https://doi.org/10.1001/jamanetworkopen.2023.16128, https://doi.org/10.1001/jamanetworkopen.2023.16128
4. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines **28**(4), 689–707 (Dec 2018). https://doi.org/10.1007/s11023-018-9482-5, http://link.springer.com/10.1007/s11023-018-9482-5
5. Forrest, K.B.: Utilitarianism versus Justice as Fairness, chap. Chapter 1, pp. 1–9 (2021). https://doi.org/10.1142/9789811232732_0001, https://www.worldscientific.com/doi/abs/10.1142/9789811232732_0001
6. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun. ACM **64**(4), 136–143 (mar 2021). https://doi.org/10.1145/3433949, https://doi.org/10.1145/3433949
7. Goldman, B., Cropanzano, R.: "justice" and "fairness" are not the same thing. Journal of Organizational Behavior **36**(2), 313–318 (2015). https://doi.org/https://doi.org/10.1002/job.1956, https://onlinelibrary.wiley.com/doi/abs/10.1002/job.1956
8. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning (2016)
9. Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L.H., Quinn, G.: Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making (2022)
10. Jones, D.A., Skarlicki, D.P.: How perceptions of fairness can change: A dynamic model of organizational justice. Organizational Psychology Review **3**(2), 138–160 (2013). https://doi.org/10.1177/2041386612461665, https://doi.org/10.1177/2041386612461665
11. Landsbergen, D., Coursey, D.H., Loveless, S., Shangraw, R.F.: Decision quality, confidence, and commitment with expert systems: An experimental study. Journal of Public Administration Research and Theory: J-PART **7**(1), 131–157 (1997), http://www.jstor.org/stable/1181549
12. Lockey, S., Gillespie, N., Holm, D., Asadi Someh, I.: A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions (01 2021). https://doi.org/10.24251/HICSS.2021.664
13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (jul 2021). https://doi.org/10.1145/3457607, https://doi.org/10.1145/3457607
14. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 107–118. PMLR (23–24 Feb 2018), https://proceedings.mlr.press/v81/menon18a.html
15. Moon, K., Blackman, D.: A Guide to Understanding Social Science Research for Natural Scientists: Social Science for Natural Scientists. Conservation Biology **28**(5), 1167–1177 (Oct 2014). https://doi.org/10.1111/cobi.12326, https://onlinelibrary.wiley.com/doi/10.1111/cobi.12326

16. Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., Strappelli, L.: Toward involving end-users in interactive human-in-the-loop ai fairness. ACM Trans. Interact. Intell. Syst. **12**(3) (jul 2022). https://doi.org/10.1145/3514258, https://doi.org/10.1145/3514258
17. Rueda, J., Rodríguez, J.D., Jounou, I.P., Hortal-Carmona, J., Ausín, T., Rodríguez-Arias, D.: "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. AI & SOCIETY (Dec 2022). https://doi.org/10.1007/s00146-022-01614-9, https://doi.org/10.1007/s00146-022-01614-9
18. Sagheb-Tehrani, M.: Expert systems development: Some issues of design process. SIGSOFT Softw. Eng. Notes **30**(2), 1–5 (mar 2005). https://doi.org/10.1145/1050849.1050864, https://doi.org/10.1145/1050849.1050864
19. Samek, W., Müller, K.R.: Towards Explainable Artificial Intelligence. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 5–22. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_1, https://doi.org/10.1007/978-3-030-28954-6_1
20. Srivastava, M., Heidari, H., Krause, A.: Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 2459–2468. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330664, https://doi.org/10.1145/3292500.3330664
21. Swartout, W.R., Moore, J.D.: Explanation in second generation expert systems. In: David, J.M., Krivine, J.P., Simmons, R. (eds.) Second Generation Expert Systems. pp. 543–585. Springer Berlin Heidelberg, Berlin, Heidelberg (1993). https://doi.org/10.1007/978-3-642-77927-5_24
22. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems **32**(11), 4793–4813 (2021). https://doi.org/10.1109/TNNLS.2020.3027314
23. Wang, N.: "black box justice": Robot judges and ai-based judgment processes in china's court system. In: 2020 IEEE International Symposium on Technology and Society (ISTAS). pp. 58–65 (2020). https://doi.org/10.1109/ISTAS50296.2020.9462216
24. Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–12. CHI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300509, https://doi.org/10.1145/3290605.3300509
25. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 295–305. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372852, https://doi.org/10.1145/3351095.3372852