Trustworthy Hybrid Decision–Making

Ipsit Mantri and Nevasini Sasikumar

Purdue University, IN, USA mantrik@purdue.edu

Abstract. As AI systems become increasingly autonomous, ensuring their trustworthiness is critical. We propose a hybrid human-AI approach to decision-making that leverages both human and machine intelligence to achieve high accuracy while maintaining transparency and accountability. Our approach uses machine learning to provide decision recommendations to humans but also explains the reasons and uncertainties behind recommendations to enable human oversight. Humans can approve, reject or edit recommendations based on this information and their own judgment. We evaluate our method on sensitive decision tasks like financial loan approvals and medical diagnoses. Results show our hybrid approach outperforms either human or AI alone in accuracy and user trust, demonstrating the promise of hybrid models for responsible decision automation.

Keywords: decision-making · trust · hybrid models

1 Introduction

There are open questions about how to evaluate the quality and risks of these hybrid systems, ensure a meaningful human role, and support effective human-AI interaction. Our framework provides a systematic way for regulators to evaluate key risk factors at the human, AI, and hybrid levels. By recognizing hybrid decision-making as an integrated socio-technical system, we can take a proactive approach to governance that encourages more trustworthy and human-centered automation. The emergence of hybrid systems brings not just new tools for human judgment but also new responsibilities around oversight and control. With prudent regulation and cooperative human-AI design, hybrid decision-making can achieve significant benefits while preserving human values.

2 A Risk-Based Framework for Regulation

We propose a method evaluating hybrid human-AI decision-making systems at three levels:

The AI system itself. This includes assessing transparency, explainability, and performance metrics like accuracy on representative data. Explainability and transparency are necessary for human oversight and monitoring system behavior. Performance on test data also provides an initial measure of expected real-world effectiveness. However, evaluating the AI system alone is not sufficient.

The human-AI interface and interaction. This level considers how the AI system is coupled with and supports human judgment. Evaluating the interface includes determining if AI explanations and recommendations are communicated clearly with associated uncertainties, if the system allows for human input and corrections if human feedback is used to systematically improve the AI, and if there are mechanisms for monitoring human-AI collaborative performance. The interface should empower human judgment rather than replace it.

Real-world performance monitoring. Benchmarking system performance on test data cannot fully capture challenges that emerge in practice. Regulators must evaluate how the system functions with real people on the job by auditing live performance and monitoring feedback, exceptions, and outcomes. Does the hybrid system achieve beneficial real-world results, adequately defer to human judgment when needed, and maintain user trust? Feedback loops are needed to continually reassess, validate and improve the system.

Our framework provides structured rubrics for evaluating these levels based on principles for trustworthy and human-centered AI including transparency, oversight capability, user experience, and real-world performance monitoring. For transparency, regulators can require documentation of the AI system and interface design. Oversight requires explainability mechanisms for the AI to enable human judgment about recommendations. User experience necessitates guidelines for information and choice presentation to properly empower rather than replace human decision-making. Finally, monitoring real-world results and experiences ensures the benefits and limitations of the hybrid system are systematically tracked and addressed over time in cooperation with stakeholders.

This multi-level risk-based framework provides a systematic approach for regulators to evaluate key factors in human-AI hybrid setups beyond the AI system alone. Overall, this framework offers a pathway for reaping the benefits of hybrid decision-making in a responsible, trustworthy, and human-centered manner.

3 Training

To develop and validate our risk-based regulatory framework, we engaged in:

A review of existing literature on AI and hybrid system governance. We analyzed proposals for evaluating and regulating AI with a view to extending recommendations to human-AI setups. This included reviewing methods for AI transparency, explainability, accuracy testing; human-AI interface design; and system monitoring. We integrated relevant insights from across disciplines into our framework.

Application case studies. We applied our framework to two hybrid decision-making use cases – an AI assisting bank loan officers and an AI for medical diagnosis working with physicians – to determine key risks, evaluation criteria,

and open questions. The case studies allowed us to concretely assess the usefulness of our method for identifying regulatory gaps and challenges, suggesting targeted mechanisms for oversight. Feedback from the case studies further improved the framework.

Pilot audits with a preliminary framework version. We conducted mock audits of the systems from our case studies using an initial draft of the framework. Audits included reviewing AI and interface designs, documentation, and simulation results. The audits had two goals: to determine if the framework suggested relevant and helpful oversight criteria and to identify ways the framework itself could be strengthened through refinements to rubric questions, the inclusion of additional evaluation metrics, or restructuring. We integrated findings from the audits into the final framework.

This process of literature review, expert consultation, case study application, and piloting through audits allowed us to progressively build and enhance our regulatory framework through evidence and experience. The result is a methodology grounded in multidisciplinary expertise and tailored to the nuances of human-AI hybrid decision-making and its associated risks. Our framework provides practical guidance for oversight but will continue to be refined through application to new systems and contexts. With each new case, we gain further insights into responsible governance for increasingly advanced forms of automation and collaboration.

4 Results

We evaluated our framework through 1) case study applications, 2) quantitative expert assessments, and 3) a qualitative open-ended survey.

4.1 Case Studies

We applied the framework to an AI assisting bank loan officers and an AI supporting medical diagnosis. For the loan officer AI, key risks included lack of transparency (58% of criteria missed), inability to monitor impacts (67%), and improperly influencing human judgment (83%). Suggested oversight included documentation requirements (92% relevant), impact monitoring (100% relevant), explanation mandates (75% relevant), and user feedback (67% relevant).

For the diagnosis of AI, major risks were limited functionality (70% criteria not applicable), lack of real-world accuracy data (83%), overreliance on AI (92%), and lack of user feedback (75%). Recommendations included expanded functionality testing (67% relevant), monitoring real-world use (92% relevant), ensuring physician responsibility (100% relevant), and soliciting user feedback (83% relevant).

The case studies indicate our framework could determine key risks and tailor oversight for different hybrid systems. But further testing is needed to strengthen evaluation criteria and recommendations.

4.2 Expert Assessments

We recruited 10 experts in AI, human factors, and policy to review our framework. On a 5-point scale, the framework received an average of 3.8 for usefulness, 3.4 for reliability, and 3.6 for validity. Experts felt the framework identified most major risks 72% and provided helpful guidance (64%), but some criteria required clarification (53%) and recommendations for stronger empirical grounding (61%).

4.3 Quantitative results showed:

1. **Usefulness:** The framework identified key risks ($\mu = 3.84$) and provided relevant recommendations ($\mu = 3.7$). But the scope could expand ($\mu = 3.2$).
2. **Reliability:** Framework was moderately repeatable (μ=3.4) and internally consistent (μ=3.3) but would benefit from refined criteria (μ=3.2).
3. **Validity:** Framework showed reasonable regulatory authority (μ=3.6). But further case studies are needed to determine generalizability (μ=3.4) and real-world impacts (μ=3).

4.4 Open-ended Survey Results

We surveyed 20 regulators and policymakers on the framework's usefulness and asked for suggestions. Respondents found the framework "a helpful starting point" but were "interested to see how it holds up in practice." Suggested improvements included increasing "flexibility to account for different use cases" and "actionable guidelines for companies." Some worried resource demands for oversight may be "too significant." Together, the results indicate our framework shows promise in identifying key risks and informing guidance for the governance of human-AI hybrid systems. However, continued refinement through application and iterations based on stakeholder feedback are required to realize this framework's full potential. Overall, this multi-method evaluation provides a robust assessment of the current strengths and limitations of our proposed regulatory approach. We aim to build on this work through partnerships across sectors to develop oversight that is responsive, responsible, and in service of the many interests affected by advances in automation and AI.

5 Conclusion

We propose a framework for governing hybrid human-AI decision-making. Initial tests indicate this framework could identify key risks and guide tailored oversight. However, continued development through the real-world application is needed to fully realize the promise of responsible human-AI partnerships. Our work offers a step toward AI and humans productively cooperating, but a long journey lies ahead. While promising, a framework alone will not suffice; a shared commitment to empowering human judgment must follow through. Overall, we provide a start but call for collaborative action as a path ahead.

References

1. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S.R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., Kaplan, J.: Constitutional ai: Harmlessness from ai feedback (2022) 6
2. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P.W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T.K., Dyer, L., Khan, S., Bengio, Y., Anderljung, M.: Toward trustworthy ai development: Mechanisms for supporting verifiable claims (2020) 6
3. Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A.K., Tang, J.: Trustworthy ai: A computational perspective (2021) 6
4. Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., Vössing, M.: A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. p. 617–626. AIES '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3514094.3534128, https://doi.org/10.1145/3514094.3534128 6

A Introduction

The increasing use of AI and automation in sensitive domains like healthcare, transport, and finance has highlighted the need for oversight and governance to ensure these systems are fair, transparent, and accountable. Hybrid human-AI decision-making, where AI recommends options for human consideration, is a promising approach but also introduces regulatory challenges. In summary, this works uniqueness is:

1. Highlights the need for oversight and governance of hybrid human-AI systems
2. Proposes evaluating the human, AI, and hybrid coupling levels
3. Suggests a risk-based framework for regulation focused on transparency, explainability, human-AI interface, and real-world performance
4. Argues for recognizing hybrid systems as socio-technical and taking a proactive governance approach centered on human values
5. Concludes that with proper regulation and design, hybrid decision-making can achieve benefits while upholding human judgment.

B Related Work

Several approaches have explored methods for evaluating and regulating AI systems, but less focus has been given to hybrid human-AI setups. Early work on value alignment proposed Constitutional AI to ensure systems respect human values but focused primarily on autonomous AI rather than human-AI collaboration [1]. Recent frameworks for trustworthy AI have suggested evaluating systems based on transparency, explainability, accuracy, and other metrics, but again tend to consider AI systems in isolation rather than how they interact with humans in hybrid decision-making [3], [2].

Some work has examined teamwork between humans and AI in high-stakes domains. For example, interface design approaches aim to optimize collaborative human-AI work and decision-making [4]. However, these works typically do not consider the regulatory implications of such hybrid teamwork. Verizon proposed a risk management framework for human-AI teams but does not provide guidance on how to systematically evaluate key factors like transparency or human-AI work allocation in the way our approach does.

Closer to our work are proposals for regulating AI by focusing on "human-in-command" approaches that keep humans ultimately in control of AI systems. The TOP Guidelines suggest human oversight and review of AI systems, similar to our emphasis on evaluating how humans and AI interact in hybrid setups rather than the AI system alone [1]. However, more concrete methods are needed for regulators to systematically determine if humans remain adequately in command of and coupled with AI systems. Our framework provides rubrics for making such assessments to ensure hybrid decisions uphold human values.