Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026465472

A Review on Human-Centered Machine Learning for Decision Support in High-Stakes Domains

Yuhao Li*

Faculty of Science, UNSW Sydney, Sydney, NSW 2052, Australia

**Corresponding author: Yuhao Li.*

Abstract

Machine learning is increasingly used to support important decisions in healthcare, public services, finance, and aviation. These systems can help with prediction, risk assessment, resource allocation, and professional decision-making. However, high accuracy does not always mean that a system is safe, fair, easy to understand, or suitable for real use. This paper reviews human-centered machine learning for decision support in high-stakes domains. It discusses common machine learning methods and several important human-centered requirements, including explainability, appropriate trust, usability, workflow integration, fairness, accountability, and human oversight. It also compares these requirements across different fields. Current studies show that many systems perform well in experiments or on historical data, but there is still limited evidence from real-world use. Explanations may increase trust, but they do not always improve decision quality. In some cases, they may even cause people to rely too much on wrong AI advice. Different fields also have different priorities. Healthcare focuses on patient safety, public services focus on fairness, finance focuses on rules and appeals, while aviation and manufacturing focus more on reliability, situation awareness, and skill retention. This paper argues that machine learning should support human judgment instead of replacing it.

Keywords

human-centered AI, machine learning, decision support systems, high-stakes decision-making, human–AI collaboration

1. Introduction

Nowadays, Machine Learning (ML) is being used in lots of modern decision in high-stakes domains and become an important part of modern decision support systems. In many fields, ML models are used for risk prediction, classification, action recommendation, resource allocation, and professional decision support. These applications are usually called High-Stakes Decision Settings. This is because system outputs might be affecting people's lives, health, rights, or social resources directly [1, 2]. They differ substantially from conventional recommendation systems and low-risk predictive tasks. The wrong decisions in high-stakes domains may lead to very serious consequences. Therefore, in these cases, ML systems should not only be seen as technical tools that aim for high prediction accuracy. They should also be understood as socio-technical systems embedded in real society, organizations, and professional workflows [2, 3]. Traditional ML research

often focuses on model performance on test datasets. This performance is often measured by metrics such as accuracy, AUC, sensitivity, specificity, precision, and recall. These metrics are important for evaluating whether a model has predictive ability. However, they cannot fully show whether a model is suitable for real-world deployment. Waldock et al. reviewed 50 studies on AI clinical decision support systems. They pointed out that although these systems showed some diagnostic ability, many studies were still based on retrospective validation. These studies also lacked evidence about real clinical workflows and implementation effects [4]. This shows that ML research in high-stakes domains needs to be further expansion. The research focus should not only stay on "whether the model is accurate." It should also focus on "whether the model can truly be used safely, reasonably, and responsibly." This paper focuses on "Human-Centered Machine Learning Decision Support." human-centered does not mean rejecting the value of algorithms. It emphasizes that ML systems should serve human decision-makers. These systems should help humans make safer, clearer, and more responsible decisions [5]. From this perspective, the research question is not only "how accurate the model is." Therefore, the main argument of this paper is that machine learning systems in high-stakes domains should improve human judgment. They should not simply replace human judgment [2, 3].

2. High-Stakes Decision Support and Human-Centered AI

2.1 High-Stakes Domains

Hight-stakes Domains are decision environments where wrong decisions may cause serious consequences. These consequences may include physical harm, medical accidents, financial loss, legal responsibility, social unfairness, or public safety problems. A common feature of these domains is that Decision-Making is not only a technical problem. Many High-Stakes Decisions include uncertainty, professional experience, ethical judgment, legal rules, and organizational workflows. Therefore, ML in High-Stakes Domains should not only aim for automation. Full automation is not suitable in many situations. This is because humans still need to make judgments, give explanations, and take responsibility. A more reasonable direction is to use ML as a decision support tool. ML can help humans process complex information, identify possible risks, and improve the consistency and efficiency of decisions [3, 5].

2.2 Decision Support Systems

A decision support system is a computer-based system that helps Human Decision-Makers analyse problems, process data, and produce suggestions. Early decision support systems usually relied on expert rules, statistical models, or knowledge bases. With the development of ML, modern decision support systems increasingly use data-driven methods. These systems can learn patterns from large historical datasets and generate predictions or recommendations [4, 5]. In High-Stakes Domains, it is important to distinguish "decision support" from "decision replacement." Decision support means that the system provides information, predictions, explanations, or recommendations. However, the final decision is still made by humans. Decision replacement means that the system directly makes the decision. In this case, humans have little or even no involvement. In healthcare, justice, and public services, fully replacing human decisions can create ethical and responsibility problems. Therefore, this paper mainly discusses how ML can support human decisions. This paper does not focus on how ML can fully replace human decisions [2, 3]. A good ML decision support system should not only give a prediction result. It should also help users understand the meaning of this result. For example, the system should not only tell a doctor that "this patient is High-Risk." It should also explain which factors influenced this judgment. The system should also show how uncertain the model is. In addition, the system can provide similar cases or suggestions for further tests. Only this kind of system can truly support professional judgment [6, 7].

2.3 Human-Centered AI

Human-Centered AI emphasizes that the design and deployment of AI systems should focus on humans. Here, "humans" include human needs, abilities, values, and responsibilities. For decision support systems, this means that the system should not only be effective. It should also be understandable, usable, challengeable, controllable, and accountable [5]. Human-Centered AI is not simply adding an explanation module to a model. It requires researchers and developers to consider users and the use environment from the beginning of system design. Different users need different types of explanations. Doctors may need clinical evidence and risk

factors. Financial analysts may need feature contributions and compliance explanations, while social workers prioritize family contextual information and institutional constraints. Barredo Arrieta et al. pointed out that explainability needs to consider the target audience [7]. This is because whether an explanation is useful depends on whether users can understand it. Whether an explanation is useful also depends on whether it serves a specific task and purpose [7]. A model that works well in an experimental environment may not work well in a real organization [2, 3]. Therefore, the evaluation of High-Stakes ML systems must go beyond the model itself. The evaluation should also pay attention to people, workflows and institutions.

3. Machine Learning Methods for Decision Support

In High-Stakes Decision Support, different types of machine learning methods are suitable for different tasks and data environments. These categories include traditional machine learning, convolutional neural networks, vision transformers, graph neural networks, Bayesian models, generative AI, and federated learning [1]. First, traditional machine learning models are still very important, such as logistic regression, decision trees, random forests, gradient boosting trees, and support vector machines. These methods are often used for structured data, such as electronic health records, financial transaction data, demographic data, and risk assessment tables. Their advantages are that they are relatively easier to explain, cheaper to train and deploy, and more suitable for settings that need transparency [1]. Second, deep learning models have advantages when handling complex and unstructured data. For example, convolutional neural networks are often used in medical imaging, industrial inspection, and agricultural disease detection. Transformer models are often used for text, images, and multimodal data. Deep learning can extract high-level features from complex data, so it performs well in image recognition, natural language processing, and speech signal analysis [1, 4]. However, deep learning models are usually harder to explain and may create black-box problems. In high-stakes settings, this lack of transparency can affect user trust and responsibility judgment [7]. Third, Bayesian models and uncertainty-aware models are especially important in high-stakes domains. High-stakes decisions are usually not simple “yes” or “no” decisions. Users also need to understand how confident the model is. If the model can show uncertainty, users can know when to accept the model’s advice and when to ask for human review or more evidence. Existing reviews on high-stakes AI also emphasize that uncertainty quantification is an important deployment requirement for safety-critical systems [1]. Fourth, generative AI and large language models are becoming new decision support tools. Large language models can be used to summarize medical records, generate reports, support information search, explain complex concepts, and help professionals with initial analysis. However, research on human-AI collaboration shows that LLMs may bring new collaboration risks. For example, users may think the system truly understands the problem because its language is fluent, or responsibility may become unclear in co-created work [3]. Therefore, in high-stakes domains, large language models should not be used as independent decision-makers. They should be used as supervised information support tools. Overall, the choice of machine learning method should not depend only on prediction performance. It should also consider explainability, data type, deployment environment, computing cost, user needs, and responsibility requirements. In high-stakes settings, the most complex model is not always the most suitable model. A simpler but transparent, stable, and auditable model may sometimes be more suitable for real use than a black-box model with slightly higher accuracy [1, 7].

4. Human-Centered Requirements in High-Stakes Machine Learning Decision Support

4.1 Explainability and Understandability

Explainability is one of the most important requirements for high-stakes machine learning systems. Users need to understand why the system gives a recommendation, especially when the recommendation may affect life, health, freedom, or resource allocation. If the system only gives a score or label without reasons, users may find it difficult to decide whether they should accept the recommendation [7]. Explainability can help users understand model behavior. It can also help developers find model problems. For example, if a medical model mainly uses variables that are not related to the disease, explanation tools may help identify this problem. Similarly, in finance or public services, explanations can be used to check whether the model uses unfair or unsuitable features [7]. However, explainability is not a perfect solution. Haddadian et al. reviewed XAI in clinical decision support and found that explanations often increase perceived trust and acceptance, but they do not always improve decision quality. In some cases, when AI advice is wrong or biased, explanations may

even increase harmful reliance [6]. Therefore, in high-stakes settings, we should not simply ask for "more explanations." We should ask for "explanations that fit the user and the task." Good explanations should help users judge when the model is reliable and when it is not, instead of only making users more willing to accept model outputs [6, 7]. In finance, clear explanation formats and simple interface design can help users understand AI decisions better [8].

4.2 Trust and Appropriate Reliance

Trust is a key issue in machine learning decision support. However, the goal of high-stakes systems should not be to make users trust AI as much as possible. The goal should be appropriate trust. Appropriate trust means that users can use the model's advice when the model is reliable, but they can question or reject the advice when the model is unreliable, biased, or outside its valid use range [3, 6]. Both over-reliance and under-reliance are problems. Over-reliance means that users accept AI advice without enough judgment, even when the advice is wrong. Under-reliance means that users ignore useful AI advice because they do not trust the system. Dong et al. pointed out that trust calibration, reliance behavior, cognitive engagement, skill retention, accountability, and transparency are repeated human-AI collaboration challenges across different AI generations [3]. To support appropriate reliance, the system needs to provide information such as confidence, uncertainty, valid use range, and error risk. Users also need training so that they can understand the limits of the model. In aviation, too much reliance on automation may reduce pilots' attention and make it harder for them to take control in an emergency [9].

4.3 Usability and Workflow Integration

Even if a machine learning model has strong technical performance, it cannot create real value if it is difficult to use. Usability includes whether the interface is clear, whether the system is easy to operate, whether there is too much information, whether it increases users' cognitive load, and whether it can fit into existing workflows [4, 5]. Zalfani et al. showed that adding explanation mechanisms, user feedback, and human oversight into decision support systems can improve user trust, understanding of transparency, usability, and decision accuracy [5]. Therefore, when evaluating machine learning decision support systems, researchers should include user experience, task completion time, cognitive load, workflow compatibility, and real adoption rate. Reporting only model accuracy cannot show whether the system is truly useful. Waldock et al. also argued that AI clinical decision support research should move from technical performance to implementation metrics, such as clinical workflow, clinician adoption, decision time, and real clinical outcomes [4].

4.4 Fairness and Bias

Machine learning systems in high-stakes domains may inherit or even increase bias in historical data. If the training data itself reflects social inequality, the model may learn unfair patterns. For example, in healthcare, some groups may appear differently in the data because they had less access to medical resources in the past. In finance, credit data may reflect inequality related to income, area, or race. In public services, historical administrative data may be affected by institutional bias [2, 4]. Fairness is not only a technical issue. It is also an ethical and social issue. Developers need to evaluate model performance across different groups, such as gender, age, region, and socioeconomic status. The system should also monitor bias after deployment, because model performance in the real world may change over time [4]. In addition, affected people should have the chance to understand and challenge decisions involving AI. If a person loses access to a loan, welfare, or service because of an algorithmic assessment, but cannot know the reason or appeal the decision, this will harm procedural justice and trust in institutions. Therefore, high-stakes machine learning systems need to combine fairness, transparency, and appeal mechanisms [2, 4].

4.5 Accountability and Governance

Accountability is another key issue in high-stakes machine learning systems. When AI advice leads to a wrong result, who should be responsible? Is it the professional who uses the system, the engineer who builds the model, the organization that deploys the system, or the regulator? If responsibility is unclear, the system may create "responsibility diffusion," where every actor thinks the responsibility does not fully belong to them [1, 3]. Therefore, high-stakes systems need clear governance mechanisms. The system should record the model

version, input data, output results, user actions, and final decision process for later auditing. Organizations should also clearly define which decisions can be supported by AI, which decisions must be reviewed by humans, and which situations require senior review or expert consultation [1]. Human oversight should not be only a formal "human-in-the-loop." Saxena et al. argued that if we do not understand how algorithms affect human decisions and how different types of decision support affect the real possibility of human intervention, simply asking for a human-in-the-loop may not be effective [2]. Therefore, truly effective human oversight needs institutional support, including training, authority, explanation tools, review procedures, and responsibility rules.

5. Applications in High-Stakes Domains

5.1 Healthcare

Healthcare is one of the most active application areas for machine learning decision support. AI clinical decision support systems have been used for diagnosis, prognosis prediction, triage, treatment recommendation, antibiotic selection, sepsis prediction, and medical image analysis. Machine learning can process large amounts of complex clinical data and help doctors find risk patterns, so it can improve decision efficiency and consistency [4]. However, healthcare also clearly shows the gap between technical performance and real clinical value. Waldock et al. analyzed 50 studies and found that AI-CDSS showed some diagnostic performance, but only some studies involved prospective deployment. Many studies mainly reported technical metrics and lacked clinical workflow data [4]. This findings means that good model performance on retrospective data does not necessarily mean that the model can truly improve doctors' decisions or patient outcomes. Healthcare decision support especially needs explainability. Doctors need not only to make decisions, but also to explain their decisions to patients, colleagues, and legal institutions. If AI recommendations cannot be explained, doctors may not want to use the system, or they may find it hard to take responsibility after using it. However, explanations themselves may also bring risks. A clinical XAI review found that explanations may increase trust, but they do not always improve decision correctness. They may even increase over-reliance when AI is wrong [6]. Therefore, the focus of healthcare AI should not only be providing explanations. It should help doctors develop appropriate reliance. Future healthcare AI decision support should pay more attention to real clinical deployment, including workflow integration, clinician behavior, patient outcomes, fairness, and responsibility mechanisms. Model performance is still important, but it is only one part of clinical value [4, 5].

5.2 Public Sector and Child Welfare

The public sector is also an important area for high-stakes machine learning applications. Algorithmic tools are used in welfare allocation, criminal justice, education, housing, employment services, and child protection. These systems usually aim to improve efficiency, save resources, and reduce inconsistency in human decisions. However, public sector decisions often involve complex social contexts and value judgments. They cannot simply be turned into prediction problems [2]. Child welfare is a typical example. In child protection systems, algorithms may be used to assess the risk of harm to children, judge service needs, or support placement decisions. These tools may help workers identify high-risk cases, but they may also make workers focus too much on risk scores and ignore family relationships, social support, and institutional factors. Through long-term fieldwork in a child welfare agency, Saxena et al. argued that public sector algorithmic systems need to support existing bureaucratic processes and strengthen human discretion, rather than replace it [2]. Machine learning systems in the public sector may also affect professional discretion. Social workers and public service workers usually need to make judgments based on specific contexts. If an algorithmic tool is too dominant, workers may directly accept algorithmic advice because of time pressure, responsibility pressure, or lack of experience. This may weaken professional judgment and make decision-making too mechanical [2]. Therefore, machine learning systems in the public sector should be designed to strengthen human discretion, not replace it. The system should help workers organize information, identify problems, and contemplate different options, instead of only giving a single risk score. Algorithmic decision support in the public sector also needs to consider law, policy, organizational resources, and citizens' rights [2].

5.3 Finance

Machine learning is widely used in finance. It can support credit checks, fraud detection, risk analysis, and loan decisions. These systems can process large amounts of financial data. They can also help analysts make faster and more consistent decisions. However, financial decisions may affect whether people can obtain loans or other financial services. Therefore, users need clear and useful explanations. [8] studied different types of AI explanations in financial decision-making. They found that explanation format and interface design can affect how well users understand the system. Simple explanations that match the task are often easier to use. This shows that financial AI systems should not only give correct technical information. They should also help analysts and customers understand and question important decisions. However, more real-world studies are still needed to test whether these explanations improve decision quality and reduce over-reliance.

5.4 Aviation

Aviation is a high-stakes field. AI and automation are used for flight control, monitoring, navigation, and risk detection. Automation can improve flight performance and reduce pilot workload. However, too much automation may reduce human attention and situation awareness. [9] studied professional pilots under different levels of automation in a flight simulator. They found that higher automation improved flight performance and reduced mental workload. However, it also changed how pilots looked at flight instruments and reduced their attention to some important information. This is a serious human-centered problem because pilots must be ready to notice problems and take control when automation fails. Therefore, aviation systems should balance automation with human attention, clear control transfer, and skill retention. More research is also needed on the long-term effects of automation on pilot skills and emergency response.

5.5 Cross-Domain Comparison

Although these domains share several common problems, their main human-centered needs are different. Healthcare focuses on patient safety, clinical workflow, and appropriate reliance. Public services focus more on fairness, professional discretion, and citizens' rights. Finance requires clear explanations, compliance, and appeal processes. Aviation focuses on situation awareness, skill retention, and safe control transfer. These differences show that one general design or evaluation method may not work well in every high-stakes domain.

6. Current Challenges and Research Gaps

Based on the existing literature, high-stakes machine learning decision support faces several important challenges. First, there is a gap between model performance and real-world deployment. Many studies can show model accuracy on datasets, but they lack validation in real environments. A model that performs well on experimental data does not necessarily fit into workflows, become accepted by users, or improve real outcomes [4]. Second, there is a gap between explanation and decision quality. Explanations can improve transparency and trust, but they do not always improve decision correctness. If explanations are poorly designed, they may even increase users' trust in wrong advice. Therefore, future research should focus on how explanations affect user behavior, not only on whether explanations exist [6, 7]. Third, there is a gap between fairness and governance. Many studies recognize that models may have bias, but practical mechanisms for bias monitoring, auditing, appeals, and responsibility are still not well developed. Fairness should not only be considered during model training. It should be included in the whole process of data collection, model development, deployment, and continuous monitoring [1, 4]. Fourth, there is a problem of formal human oversight. Many systems claim to use a human-in-the-loop design, but it is still unclear whether humans really have the ability and authority to intervene. If users only passively confirm AI outputs, human oversight may become a way to transfer responsibility, not a real safety mechanism [2]. Fifth, different high-stakes domains have clear differences. Healthcare, finance, public services, and safety systems have different goals, data, users, and responsibility structures. Therefore, one general solution cannot solve all problems. Future research needs to develop domain-sensitive evaluation frameworks [1, 3].

7. Future Research Directions

Future research should move from a model-centered view to a human-centered and deployment-centered view. First, studies should not only report accuracy, AUC, sensitivity, and specificity. They should also report workflow integration, user adoption, decision quality, cognitive load, fairness, and real outcomes. Especially in healthcare and public services, real-world usefulness is more important than test dataset performance alone [4]. Second, future research should develop metrics for "appropriate reliance." Researchers need to distinguish whether users accept AI advice when AI is correct and whether users reject AI advice when AI is wrong. These metrics are more useful than simply measuring how much users trust AI [3, 6]. Third, when possible, researchers should consider more interpretable models or models with built-in explainability. For some high-stakes tasks, a slightly simpler but transparent model may be more suitable for deployment than a complex black-box model. Of course, when complex models must be used, they should be combined with uncertainty estimation, explanation tools, and auditing mechanisms [1, 7]. Fourth, system design should use participatory methods. Doctors, nurses, social workers, financial analysts, patients, citizens, regulators, and other stakeholders should take part in the design process. This can ensure that the system solves real problems instead of only optimizing technical metrics [2, 5]. Fifth, governance mechanisms need to be strengthened. High-stakes machine learning systems should include documentation, model monitoring, bias detection, audit trails, human review, and responsibility allocation. Especially for generative AI and large language models, future research needs to focus on hallucination, misinformation, responsibility, and over-trust [1, 3]. Finally, future research should pay more attention to knowledge transfer across domains. Healthcare, aviation, public services, and manufacturing have different tasks, but they all face common problems such as trust calibration, over-reliance, transparency, skill loss, and responsibility allocation. Cross-domain reviews can help researchers find common problems and avoid repeating the same mistakes in different fields [3]. The system design can also include cognitive reminders. For example, users may be asked to make an independent judgment before seeing AI advice, or the system may remind users to ask for human review when the model is uncertain [3, 6].

8. Conclusion

Machine learning has great potential in decision support for high-stakes domains. It can help professionals process complex data, find risk patterns, improve decision efficiency, and reduce inconsistency in human judgment to some extent. However, high-stakes machine learning systems cannot be evaluated only by prediction performance. Accuracy is necessary, but it is not enough [1, 4]. This paper argues that human-centered requirements should become core evaluation standards for high-stakes machine learning decision support. Explainability, trust calibration, usability, fairness, accountability, and human oversight should not be seen as extra functions. They should be considered basic requirements for system design and deployment [3, 5, 7]. Research across healthcare, public services, finance, and aviation shows that machine learning creates value only when it supports human judgment, fits real workflows, and includes clear responsibility mechanisms. However, different domains have different priorities. Financial systems need clear explanations and meaningful review, while aviation systems need continuous attention, situation awareness, skill retention, and safe human takeover.

References

[1] Human–AI Collaboration Across Decision Support, Autonomous Systems, and LLM Agents: A Systematic Review and Collaboration Convergence Framework. https://www.mdpi.com/2071-1050/18/11/5313

[2] Artificial Intelligence for High-Stakes Decision Support: Architectures, Applications, and Deployment Challenges. https://al-kindipublishers.org/index.php/fcsai/article/view/12751

[3] Explainable AI in Clinician-Facing Clinical Decision Support: A Critical Systematic Review and Evidence Map of Human-Centered Evaluations. https://www.isjtrend.com/article_243223.html

[4] Human-Centered AI for Decision Support Systems: Enhancing Usability and Trustworthiness. https://www.mdpi.com/2079-8954/14/6/651

[5] A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. https://dl.acm.org/doi/abs/10.1145/3476089

[6] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103

[7] Performance of predictive AI-based clinical decision support systems across clinical domains: A systematic review and meta-analysis. https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0001310

[8] Designing effective explainable AI: a human-centered evaluation of explanation formats in financial decision-making. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2026.1668029/full

[9] Impact of automation level on airline pilots' flying performance and visual scanning strategies: A full flight simulator study. https://www.sciencedirect.com/science/article/pii/S0003687024002333

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access