Artificial Intelligence In Medicine 149 (2024) 102780

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review

Harishankar V. Subramanian [a], Casey Canfield [a,*], Daniel B. Shank [b]

[a] Engineering Management & Systems Engineering, Missouri University of Science and Technology, 600 W 14th Street, Rolla, MO 65409, United States of America
[b] Psychological Science, Missouri University of Science and Technology, 500 W 14th Street, Rolla, MO 65409, United States of America

ARTICLE INFO

Keywords:
Explainable AI
Stakeholder engagement
Human-centered design
Human-AI team
Trust in AI
Kidney transplant

ABSTRACT

The rise of complex AI systems in healthcare and other sectors has led to a growing area of research called Explainable AI (XAI) designed to increase transparency. In this area, quantitative and qualitative studies focus on improving user trust and task performance by providing system- and prediction-level XAI features. We analyze stakeholder engagement events (interviews and workshops) on the use of AI for kidney transplantation. From this we identify themes which we use to frame a scoping literature review on current XAI features. The stakeholder engagement process lasted over nine months covering three stakeholder group's workflows, determining where AI could intervene and assessing a mock XAI decision support system. Based on the stakeholder engagement, we identify four major themes relevant to designing XAI systems – 1) use of AI predictions, 2) information included in AI predictions, 3) personalization of AI predictions for individual differences, and 4) customizing AI predictions for specific cases. Using these themes, our scoping literature review finds that providing AI predictions before, during, or after decision-making could be beneficial depending on the complexity of the stakeholder's task. Additionally, expert stakeholders like surgeons prefer minimal to no XAI features, AI prediction, and uncertainty estimates for easy use cases. However, almost all stakeholders prefer to have optional XAI features to review when needed, especially in hard-to-predict cases. The literature also suggests that providing both system- and prediction-level information is necessary to build the user's mental model of the system appropriately. Although XAI features improve users' trust in the system, human-AI team performance is not always enhanced. Overall, stakeholders prefer to have agency over the XAI interface to control the level of information based on their needs and task complexity. We conclude with suggestions for future research, especially on customizing XAI features based on preferences and tasks.

1. Introduction

Artificial intelligence (AI) technology is rapidly accelerating, opening new opportunities to integrate it into high-stakes domains such as healthcare, defense, and legal applications. However, several high-profile attempts to integrate AI into these work systems have failed. For example, using IBM's AI Watson for oncology revealed that it provided incorrect treatment recommendations [1]. Similarly, an investigation of AI in the legal domain revealed inherent racial bias in the system [2–4]. To appropriately trust AI systems, users need to be aware of an AI system's abilities and limitations.

Stakeholder engagement can support efforts to design trustworthy systems for specific applications based on stakeholders having in-depth knowledge of their own needs [5–7]. In particular, AI designers engaging with stakeholders early in the process may improve stakeholders' trust in the research and support mutual learning of each other's goals [8–10]. Additionally, stakeholder's involvement in the design process may also help identify important directions for research [11], research questions, and areas for intervention [8]. Furthermore, stakeholders' involvement in the design process can ultimately promote adoption and transparency [12]. A stakeholder engagement process typically involves drafting user workflow to identify areas of improvement and iteratively improving interfaces [13], [14]. To this end, the present paper demonstrates this process in the context of the kidney transplant placement process leading to a stakeholder-informed scoping review, where themes that emerged from stakeholder engagement work are contextualized in the context of the explainable AI (XAI) literature to identify areas for future research.

* Corresponding author.
E-mail address: canfieldci@mst.edu (C. Canfield).

https://doi.org/10.1016/j.artmed.2024.102780
Received 26 July 2023; Received in revised form 20 December 2023; Accepted 14 January 2024
Available online 20 January 2024

1.1. Explainable AI

AI and machine learning methods are typically black boxes that make it difficult for users to understand how the system works or arrives at a particular prediction. For example, deep neural networks combine multiple layers of neural networks to achieve high prediction accuracy, but this results in a complex, often non-linear structure unsuited to simple explanations [15], [16]. As a result, XAI is an expanding research area focused on making black-box models transparent through background information [17] and post-hoc explanations derived from complementary models [18]. Explainability information helps users effectively trust and use the system [19] by helping make the purpose of the system and internal functions clear. This transparency makes the system interpretable, which is required for achieving understandability [20], [21].

Explainability is a broad term that refers to providing information about the AI system. This can include system-level information (inputs and prediction patterns), prediction-level information (reasoning for a specific outcome), and model incompleteness (model's training boundary conditions) [22]. System-level (or global-level) information reveals system's operations as a function of all predictions or outcomes [17]. The goal of system-level information – which includes summary statistics, training or onboarding, and disclosures – is to help users develop a mental model of how the AI works and fits into their decision-making process [23]. This helps users understand the AI's limits [13], [24], [25]. Prediction or local-level information provides details on a specific AI prediction – how the input data maps to the output [17]. This is distinct from uncertainty information, which refers to the AI system's lack of knowledge about an outcome of interest [26], [27]. In a classification task, uncertainty information can be represented as the predicted probability of the AI's outcome matching the ground truth [26].

1.2. Kidney transplant placement process

In 2022, over 25,000 kidneys were transplanted in the United States, but the demand for donated organs outpaces the supply, with 100,000 people remaining on the kidney transplant waiting list nationally [28]. Transplantation provides recipients suffering from end-stage kidney disease with a better quality of life and long-term survival. Even with less desirable organs, transplantation is cost-effective, often cost-saving [29], and provides survival benefits to some recipients [30], [31]. Recent studies suggest substantial untapped potential for kidney utilization in the United States compared to other countries, primarily from the broader use of organs from older donors with more comorbidities [32]. In the U.S., approximately 20 % of procured deceased donor kidneys are not utilized for both avoidable and unavoidable reasons [33]. The non-utilization rate rises exponentially with measures of lower organ quality, such as higher Kidney Donor Profile Index (KDPI) scores. While some non-utilization may be medically appropriate, other cases likely reflect missed opportunities caused by delays in placing a given organ with an accepting transplant center.

The kidney transplant process includes three stakeholder groups – (1) transplant centers, (2) Organ Procurement Organizations (OPOs), and (3) transplant recipients (or candidates who later become recipients). Transplant center professionals include transplant surgeons, nephrologists, and transplant coordinators (nurses). These professionals are involved, to varying degrees, in accepting or declining a kidney offer from an OPO. OPO professionals include medical directors, operations directors, and procurement coordinators. They work with donor families for the donation and match each kidney with a transplant center.

For each organ donor, a procurement coordinator from the OPO is responsible for identifying a destination for each organ using a prioritized list based on need, proximity, and medical compatibility. Ideally, the OPO aims to place all organs before procurement begins. For each donor organ, the OPO coordinator determines how many transplant centers to make an offer to via the DonorNet platform. OPO efforts to contact transplant centers happen at all times of the day and night because deceased donors may become available at any hour. Data also suggest that overnight procurements are one of the main obstacles in placing a less desirable kidney [34].

For a kidney to be considered “hard-to-place,” OPO staff must exhaust the prioritized list by offering the kidney to all transplant centers within a 250-mile radius. At this point, the OPO may deviate from the prioritized list to pursue accelerated placement to avoid discarding the organ. Some OPOs have established decision rules where they engage in accelerated placement if the cold ischemic time (time since procurement) exceeds specific values. However, depending on logistical constraints for transporting the kidney to transplant centers (e.g., the time required for transport) and risk characteristics of the donor's kidney, the appropriate threshold to avoid kidney non-utilization varies.

When transplant centers receive organ offers via the DonorNet platform they have one hour to decline or provisionally accept the offer. Often, transplant centers will provisionally accept an offer to keep their options open, even if there is a low likelihood that they will ultimately accept the offer. The transplant team receives access to extensive information, including the donor's medical history, known risk factors for organ function (e.g., age, cause of death, diabetes, hypertension, Hepatitis C), and KDPI. After the OPO procures a kidney, transplant center staff can adjust their decision as more information becomes available based on patient input and compatibility. When deciding whether to perform a transplant, surgeons consider factors ranging from medical compatibility, competing offers, transplant team fatigue, and patient support systems, which can all affect the success of the transplant.

Further, transplant centers and individual surgeons vary in their ability to care for recipients with complications (e.g., recipients who get Hepatitis C infections from donors), in their preference for living donor transplantation, and in their risk level based on recent unsuccessful transplants. Ultimately, the surgeon has until the moment of transplantation to decide to decline a kidney offer. Offers declined at this stage are at the highest risk of non-utilization and challenging for OPOs to reallocate.

2. Methods

We used a participatory research framework to combine stakeholder engagement with a literature review to identify promising areas for future research [35], [36]. In this case, the stakeholder engagement focused on the kidney transplant placement process. This informed the analysis of the literature review, which was more broadly focused on XAI.

2.1. Stakeholder engagement

Over nine months (Dec 2020 – Sept 2021), we recruited transplant professionals and recipients to participate in workshops and interviews. We conducted a three-stage qualitative study to ask critical stakeholders 1) what they need from an AI, 2) how they make decisions with an AI, and 3) what additional information they need from/about an AI [37]. Data collection materials and aggregated results are available on Open Science Framework at https://osf.io/ju9x3/. We recorded all interactions for follow-up analysis.

2.1.1. Procedure

We invited each participant to participate in one individual interview and three workshops. Each online interview lasted 30-90 min, and each online workshop lasted 2 h. The present study is limited to feedback from transplant professionals during the June 2021 – Sept 2021 interactions, which include two workshops and one set of interviews. Feedback from recipients on the patient perspective will be addressed in future work. The first workshop and initial interviews focused on identifying the problem and where an AI decision support system would fit into the existing workflow [38]. The present study focuses on the second workshop and interviews where participants evaluated an AI decision support mock-up and the third workshop where participants evaluated AI interfaces.

In the second workshop, stakeholders reviewed the proposed system architecture and evaluated a mock-up of the AI decision support system. Consistent with a think-aloud protocol, where participants verbalize their thoughts and actions while performing a task [39], stakeholders provided feedback on the different functions and outcomes of a mock AI system. Participants reviewed four scenarios that represented a hit, miss, false positive, and false negative by the AI decision support system. Fig. A1 represents the 'miss' scenario. Each scenario included medical information on the donor's kidney (Fig. A1a) and a short list of potential recipients for the kidney (Fig. A1b). In small groups of three to six, professionals evaluated each offer. Transplant center stakeholders decided whether they would accept or decline the kidney for each candidate. OPO stakeholders evaluated the same donor characteristics and a likelihood of acceptance for a list of potential recipients (Fig. A1c). The OPO group decided which transplant centers to target for the offer and in which order. After the group reviewed the clinical data, they evaluated a prediction from a deep learning model trained on historical data (Fig. A1d). To increase input from transplant surgeons specifically, we converted the workshop content into an interview protocol and interviewed additional transplant surgeons. The interview protocol followed the same structure as the workshop.

We hosted a third workshop to solicit feedback on the proposed interfaces for the AI decision support system, highlighting potential XAI features. The discussions centered on what information they would like from the decision support system and how to provide it. We proposed four formats (see Fig. A2) that included cumulatively increasing amounts of information: (a) a binary prediction, (b) with a confidence rating, (c) with a list of which factors increased or decreased the prediction, and (d) with a sensitivity rating for each factor. The participants reviewed each format independently and provided input on what was relevant and missing to make an informed decision.

2.1.2. Recruitment

We recruited participants from the three stakeholder groups. Initially, we recruited transplant professionals from the transplant centers and OPOs in Missouri, Nebraska, Iowa, and Kansas. We shifted to recruiting nationally for transplant centers to increase the sample size for the interviews in July 2021. For transplant recipients, we recruited individuals active in the transplant community nationally. As summarized in Table 1, all the OPO professionals invited to attend joined at least one workshop or interview. In addition, over half of the transplant recipients and transplant center professionals participated in at least one event. Ultimately, 39 stakeholders participated in at least one engagement event.

2.1.3. Analysis

For this analysis, we analyzed 17.8 h of content. There were 13.5 h of content from the workshops, which included multiple breakout rooms that were recorded separately. We collected 7.3 h from the second workshop and 6.2 h from the third workshop. In addition, there were 4.3 h of content from interviews.

Table 1
Summary of participation in engagement activities by stakeholder group.

Stakeholder group	Decision support mockup workshop & interviews	Interface preference workshop	Response rate
	June-July 2021	September 2021	
Transplant Centers	12	4	16/27 (59 %)
OPOs	12	6	15/15 (100 %)
Recipients	7	5	8/12 (67 %)
Total	25	15	39/54 (72 %)

We used an inductive qualitative research process to identify key themes from the stakeholders [40]. After each workshop and interview, the researchers debriefed and summarized the new information gathered. This followed an iterative process of individual and group evaluation to develop a consensus on the interpretation, similar to the Ward method [41] High-level summaries were shared with the participants (see breakout room summaries in post materials on Open Science Framework). There were 13 key points raised in the second workshop and eight key points in the third workshop. The first author re-reviewed all the data to ensure the findings held. Based on this review, we consolidated the key points into nine themes (which are reflected in the subsections of the results).

After the literature review, we further reduced the number of themes to represent those which (1) emerged from and cover the stakeholder engagement findings and (2) had coverage in the scoping literature review. This type of approach, where stakeholders set the agenda, has also been applied in the context of establishing sustainability criteria for biofuels across diverse groups [42]. This process allowed us to identify whether stakeholder input was or was not consistent with the literature. Thus, themes from the stakeholder engagement influenced how we analyzed the literature review, rather than how we selected articles. Ultimately, we consolidated the findings into four major themes: 1) contextual use of AI predictions, 2) information included in AI predictions, 3) personalization of AI predictions for different groups, and 4) customizing AI predictions for specific cases.

2.2. Scoping review process

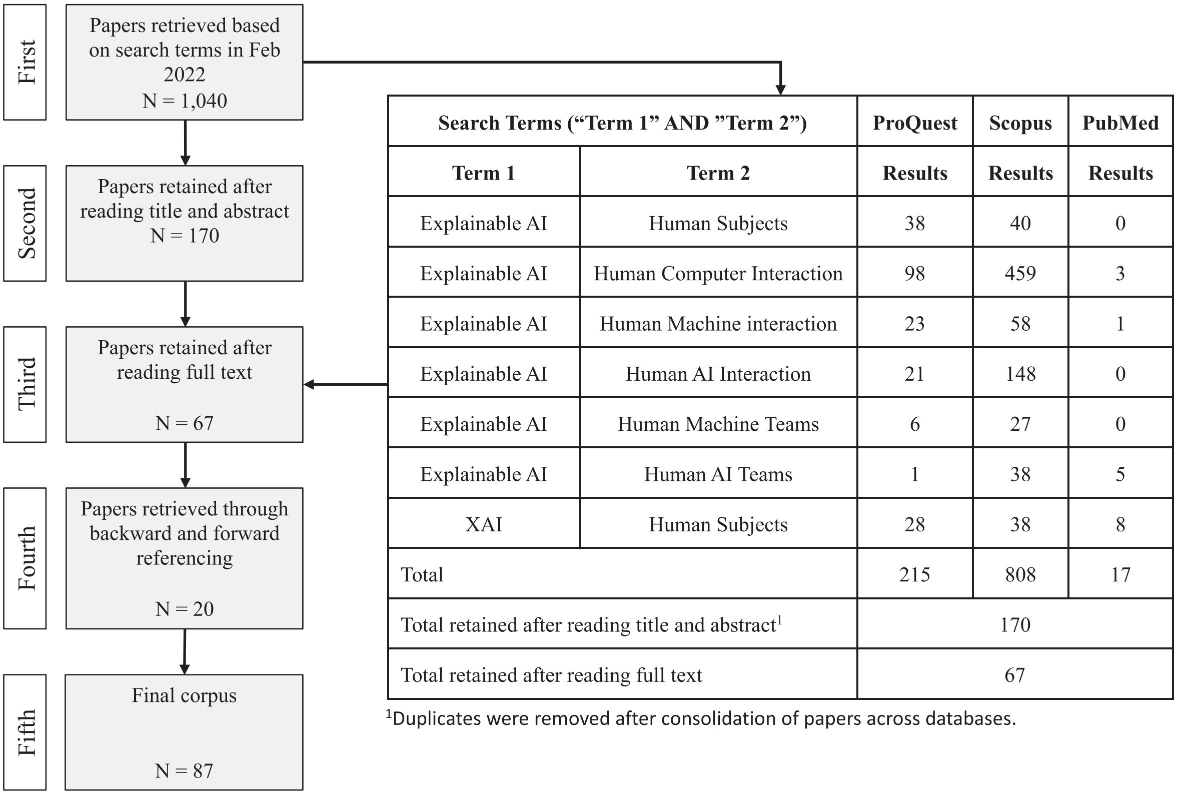
We conducted our scoping literature review following the PRISMA standards summarized in Fig. 1 [43], [44]. The results of the stakeholder engagement informed the analytic themes and organization of the literature review, rather than the search terms which were chosen to be broadly inclusive of XAI research. First, in February 2022 we searched three databases, ProQuest, Scopus, and PubMed, which generated 1040 results, including duplicates. In each database, we used the following search terms combined in pairs described in Fig. 1: Explainable AI, XAI, Human Subjects, Human-Computer/Human-Machine/Human AI Interaction, and Human-machine/Human AI teams. For example, the terms "Explainable AI" and "Human Subjects" conjoined by an "AND" were employed as Term 1 and Term 2 during the search process. These search terms were applied across three databases, utilizing the "full text" filter to retrieve articles.

Second, we reviewed the title and abstract of each paper to determine the relevance. Consequently, we excluded documents limited to the technical functionality of an AI decision support system without addressing human-AI interactions. Third, we read these 170 remaining papers for relevance and use of human-subjects research, leaving only 67 papers. Fourth, we added 20 papers based on the backward and forward references from this set of 67 articles. This final corpus of 87 articles is limited to articles published in and before 2022, with task performance and/or user trust as their outcome measure. The final corpus includes 55 quantitative, 16 qualitative, and 16 review papers which are summarized in Tables A1, A2, and A3.

This corpus was then analyzed for the themes that emerged from the stakeholder engagement process. By focusing on the themes that were brought up by the stakeholders, we identified areas that may merit future research. As described above, the nine initial themes were consolidated into the 4 reported here.

3. Stakeholder-driven literature review for transplant placement

This section provides stakeholder-driven and literature-supported


First
Papers retrieved based on search terms in Feb 2022
N = 1,040
Second
Papers retained after reading title and abstract
N = 170
Third
Papers retained after reading full text
N = 67
Fourth
Papers retrieved through backward and forward referencing
N = 20
Fifth
Final corpus
N = 87

Search Terms (“Term 1” AND ”Term 2”)		ProQuest	Scopus	PubMed
Term 1	Term 2	Results	Results	Results
Explainable AI	Human Subjects	38	40	0
Explainable AI	Human Computer Interaction	98	459	3
Explainable AI	Human Machine interaction	23	58	1
Explainable AI	Human AI Interaction	21	148	0
Explainable AI	Human Machine Teams	6	27	0
Explainable AI	Human AI Teams	1	38	5
XAI	Human Subjects	28	38	8
Total		215	808	17
Total retained after reading title and abstract[1]		170		
Total retained after reading full text		67		

[1] Duplicates were removed after consolidation of papers across databases.

Fig. 1. Literature search results following the PRISMA standards.

insights based on four critical themes – 1) contextual use of AI predictions, 2) information included in AI predictions, 3) personalization of AI predictions for different groups, and 4) customizing AI predictions for specific cases.

3.1. *Contextual use of AI predictions*

The initial design and scope of an AI tool should be dictated by the desired purpose [13]. The stakeholder engagement showed the importance of understanding context including (a) when – and for what purpose – to integrate an AI in an existing workflow and (b) what data to include in the AI predictions.

3.1.1. *Timing of AI predictions*

AI decision support systems can share their predictions before, during, and after users have reviewed the data [45]. In large part, this depends on first choosing the role of the AI, which can be integrated into the workflow as a (1) screening tool, (2) alert system, or (3) second opinion [46]. The appropriate role for the AI depends on the decision context, which is likely to vary across cases and users. In addition, different elements of the AI information (i.e., prediction, explanation) can be shared at different times.

In kidney transplants, OPOs primarily framed the AI as a useful screening tool for determining whether a kidney is “hard-to-place.” The OPO stakeholders emphasized that an immediate intervention could be beneficial (i.e., *before* reviewing the data) due to the time-sensitive nature of their mission. Knowing earlier that a kidney has a high risk of being hard-to-place allows OPOs to follow the accelerated process to increase the chances of placement. In contrast, transplant centers saw complementary roles for AI. Some stakeholders suggested that the AI could be used as a screening tool so that surgeons can concentrate on factors the AI cannot. In addition, they saw some value in having the AI act as a highlighter for crucial information that influenced its prediction. Transplant centers primarily framed AI as a type of second opinion for deciding whether to accept or decline a kidney offer.

The literature review suggests that experts tend to want alerts for time-sensitive information [47]. For example, clinical experts from a pediatric intensive care unit (ICU) preferred to know critical information immediately [48]. In addition, studies have found explainability information is most useful during, rather than before or after, the decision-making process if there is a disagreement between the AI and the user [46], [47]. For example, users requested an explanation for an autonomous vehicle's behavior during an unexpected event rather than before or after the event occurred [49]. Similarly, explanations were most useful during high-risk situations such as collisions or emergencies [50]. Lastly, the literature also suggests that it may be beneficial for stakeholders to receive an AI prediction *after* reviewing the data to make a preliminary decision [43], [45–51]. For sentiment classification of beer and book reviews, users reported that 47 % used AI predictions as a starting point and 25 % used AI as a post-check in decision-making [52]. Providing AI information after a preliminary decision may minimize concerns about over-reliance on AI, which has liability implications.

3.1.2. *Data included in AI predictions*

The appropriate unit of analysis for AI predictions varies across stakeholder groups because they perform different tasks. Transplant centers need predictions at the candidate level to identify how the offered kidney matches their patient. In the stakeholder engagement, after reviewing the AI prediction and historical clinical decision, transplant stakeholders recognized that the AI had a somewhat limited perspective based on the data available to it.

In contrast, OPOs need predictions at a transplant surgeon or transplant center level because their behavior, rather than just a candidate's characteristics, influences whether a kidney offer will be accepted and ultimately transplanted. For example, transplant surgeons have scheduling constraints, risk preferences, and center-level policies which are not available in the transplant data but do explain acceptance practices [52–54]. However, transplant data are collected and stored at the candidate level, which is associated with a specific transplant center, but not a specific transplant surgeon. Additionally, several OPO stakeholders mentioned that the AI system would be most beneficial if it included the amount of time to transport the kidney to the transplant centers [38], [53], [55]. If an OPO coordinator gets a prediction that involves a transplant center they have not recently worked with, they need to know whether the timing logistics disqualify that option.

The literature is clear that AI should provide predictions in a manner that reduces users' effort in decision-making [46], [56–59]. By better aligning with the user's decision-making process, it is easier for human users to evaluate the quality of the AI prediction [60–64].

3.2. Information included in AI predictions

Trust in an AI system influences initial adoption as well as retention of users over time [65]. Users need both (a) system-level and (b) prediction-level information to determine when to trust that AI predictions will help improve performance [66].

3.2.1. System-level information

In the stakeholder engagement, all attendees were shown a confusion matrix for the proposed AI, which was explained in detail by the moderators (Fig. A3). A confusion matrix summarizes the performance of an AI system in terms of true positives (hits), false positives (false alarms), true negatives (correct rejections), and false negatives (misses). All stakeholder groups mentioned that knowing the accuracy and other performance measures is beneficial to understand the overall system. Participants also wanted to know additional system-level information such as the AI training dataset and boundary conditions. Providing measures beyond system accuracy may be necessary for users to build an appropriate mental model of the system [67] (see Table 2). Both the transplant center and OPO stakeholders wanted the system-level information to be embedded within the DonorNet interface to be easy to access.

Informing users of an AI's limitations can help them navigate AI predictions during decision-making [24], [66–69]. In a healthcare setting, users tend to be more concerned about an AI's reliability and accuracy rather than it's reasoning or explanations [70]. Unfortunately, users, even machine learning (ML) experts, often found confusion matrix jargon difficult and hard to interpret [71]. A confusion matrix can significantly improve users' objective and subjective understanding of an AI system if it is contextualized, visualized (e.g., in a flow chart), and explained in domain-specific terms rather than generic terms like "false positives" [72].

Table 2
Examples of system-level information for an AI system for transplant surgeons.

Category	Information	Example
Purpose	Outputs	AI predicts whether to accept or decline a kidney offer for a particular patient.
Performance	Accuracy	99 % accurate
	Rate of false positives	10 % false positives; confusion matrix (see Fig. A3)
	Rate of false negatives	5 % false negatives; confusion matrix (see Fig. A3)
		F1 score (0.851 training, 0.824 holdout)
	Deployment details	Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) – (0.844 training, 0.633 holdout)
Training	Volume of training data	Trained on 1.3 million data points.
	Description of training data	Training data are from 2016 to 2021 and include the match run, patient medical history, and donor medical history.
Operation	AI pre-processing and analysis	This AI uses a deep neural network to generate predictions. Pre-processing was performed to remove missing data and balance the training data.
	Model Information	Model ID: Kidney-TXC-v1.0 Created: 01/14/2023 Last Modified: 05/16/2023 Prediction type: Classification
Limitations	Description of boundary conditions	Not appropriate for pediatric transplants

Trust and satisfaction tend to increase for models with higher accuracy, unless users perceive the AI to be inaccurate [71], [72]. Empirical evidence suggests that the model's performance, for example seeing the model make a mistake, has more impact on user trust than many XAI features do [63], [73–76]. User trust is also significantly reduced when expectations are violated, although increasing system transparency reduces the effect [77]. In a study on the onboarding needs of pathologists, researchers found that user trust is quickly lost when the system does not perform to their "gold standard" expectations [37]. Furthermore, providing information on highly influential inputs can improve user trust in the system, especially when the model has high performance accuracy [78]. However, when system-level information is provided, users, even experts, may over-trust and adhere to AI predictions irrespective of the accuracy [79]. This suggests that explicit presentation of the system's limitations and error boundaries is critical.

3.2.2. Prediction-level information

In the stakeholder engagement, surgeons overall preferred simple prediction-level information, which was a binary prediction (accept/decline) with a numerical confidence rating (Fig. A2b). The more complicated explainability information was generally consistent with their existing mental model and therefore did not provide additional insight. In contrast, OPO coordinators and recipients tended to prefer more detailed explanations of how specific inputs influenced the prediction, likely because their decision-making is less driven by clinical factors (Fig. A2d). The OPO coordinators mentioned that the detailed information would also be beneficial as a training tool for new staff and help with system transparency.

Ultimately, prediction-level information allows users to determine whether to trust a specific prediction by identifying outlier or edge cases, verifying the prediction based on the input data, and providing information on the quality of the prediction [80] (see Table 3). The challenge is in identifying an appropriate amount of information to provide without overwhelming or distracting the user [78–80]. In the long run, this has implications for adoption, where users are less likely to use AI models they do not trust or find less useful [63], [72], [81].

Evidence suggests that users can benefit from both uncertainty and explainability information, especially when it is simple and easy to understand [27], [49], [52], [82–87]. In a general question answering task, users receiving an AI prediction with confidence information about the AI's prediction showed significantly improved user accuracy, sensitivity, and reduced false positive rate compared to receiving an AI prediction

Table 3
Examples of prediction-level information about an AI system for transplant surgeons.

Category	Information	Example
Inputs	Raw data	Table of data, see Fig. A1a.
Uncertainty or Confidence	Confidence interval	95 % CI: 0.68–0.91
	Likelihood or probability	Kidney is likely (80 %) to be declined.
Explainability (text)	Counterfactual	Shows inputs for opposite prediction, see Fig. A4.
	Nearest Neighbors	Shows predictions for different similar inputs, see Fig. A5.
Explainability (visual)	Outlier indicator	Patient age is below 18, outside of the AI training boundary.
	Feature importance or contribution	Plot ranking influence of inputs, see Fig. A6.
	Heatmaps	Color coded plot showing influence of inputs, see Fig. A7.
Explainability (numerical)	Sensitivity	Measure of how much changing an input affects the output, see Fig. A2d.

alone [88]. Additionally, showing uncertainty information has also empirically been found to improve user trust in the system [27]. Research on identifying skin diseases found that providing multiple recommendations with probability metrics was significantly better at improving accuracy than a single binary prediction with probability [89], similar to the findings in an image recognition task [90]. In another image recognition study, users who received nearest neighbors explainability information performed better than those who received classification trees because it was easier to understand [91], [92]. Similarly, in another image recognition task, users receiving counterfactuals had significantly higher justified trust than users with other XAI features [93]. In a classifier task, users' accuracy improved only slightly with XAI but significantly improved when system-level model accuracy information was also provided [91], [92]. This suggests that users cannot build a full mental model based on prediction-level XAI information alone. The effectiveness of explainability methods in the quantitative literature is summarized in Table A1.

In general, it is challenging to achieve complementary performance, where the human-AI team outperforms the human or the AI alone. The goal is to achieve appropriate or justified trust, where human users can navigate when it is appropriate to trust an AI and when they have more knowledge than an AI. As a result, XAI features need to be informative, rather than convincing [94], [95]. Some studies have demonstrated complementary performance, such as in manufacturing defect identification [96] and sentiment classification of beer and book reviews [52], but there was no additional benefit associated with explainability information. In some cases, human-AI teams perform worse than the human alone because the humans are unable to identify incorrect AI recommendations [52]. For example, users may be unable to identify whether they have exceeded the context within which the AI is trained, and therefore should not trust the AI recommendations [97]. In other cases, expert users are able to perform better than an AI alone [98]. Explainability information generated by the AI lacks the ability to identify and reason why its process failed [25]. As a result, there may be value in designing communications that highlight when to be skeptical of an AI, rather than just provide explanations.

3.3. *Personalization of AI predictions for different groups*

Personalizing an AI system may be valuable to increase trust, performance, and efficiency. Therefore, the appropriate interface for an AI system will vary by users, who differ in terms of (a) expertise and (b) decision-making threshold.

3.3.1. *Expertise*

In the stakeholder engagement, the transplant professionals varied in terms of preferences, which may be driven by differences in experience, trust in AI, and comfort level with technology. In other words, preferences may vary based on both expertise in the domain and in AI systems and these can change over time as individuals gain expertise and experience [13]. When discussing the different levels of XAI information, transplant surgeons and OPO coordinators saw benefit in having more XAI features as expandable options that they can access for difficult use-cases. Additionally, one OPO coordinator stated that their trust in AI would improve if the AI provided information on how each input affected the prediction. Expert users want to be able to use their judgement to decide how to leverage AI predictions.

The literature suggests that one of the biggest differences between users of XAI is their domain expertise [99]. Novice users typically benefit from more explanations because they may not have a strong mental model for the task or the AI [97], [98]. In particular, novice users benefit from combining text and visual explainability information [56], [99] and receiving counterfactual explainability [60], [100–104]. Novice users trusted XAI more when assessing migraines, where they had high domain knowledge, compared to assessing temporal arthritis, where they had low domain knowledge [105].

Experts can effectively synthesize system-level and prediction-level AI explanations. Pathologists in cancer diagnosis desired more system-level information than prediction-level explanations [37]. In a qualitative study, radiologists wanted more prediction-level explanations whereas physicians wanted more system-level explanations [106]. In a qualitative study with experts tasked to identify a criminal suspect and motives, users found that system transparency enabled them to inspect and verify the system operation. Providing more information did not increase the cognitive load as the users with explanations actually performed the task more efficiently than users without explanations [107].

For expert users, explainability information may be more effective when it is interactive. For example, in a medical notes annotation task, an AI recommender system either provided multiple recommendations for annotating a highlighted word (interactive) or pre-annotated the highlighted word (non-interactive) [108]. Although experts were able to effectively evaluate the AI in both conditions, pre-annotations caused a loss of agency and a decrease in engagement despite their subjective reports that pre-annotations increased engagement.

In addition, users may vary in terms of AI expertise or data literacy. Users with more ML experience tend to have higher performance and are better able to critically analyze explainability features [109]. However, for users with less AI expertise, perceived understanding can decrease when asked to explain in detail how the model makes its prediction [109], [110]. Similarly, users' perceived understanding decreased when users reviewed their performance in a forward simulation task, which involves users predicting the AI's outcome [100]. Text information on how to interpret the explanations may be more effective for helping users understand system behavior and develop a mental model of the AI [64], [106].

3.3.2. *Decision-making process*

Users may also vary in terms of their decision-making process and mental model of the task itself. In the stakeholder engagement, transplant surgeons discussed how there is variation in terms of risk posture, where some surgeons are willing to transplant marginal kidneys whereas others are more conservative. OPOs also observe this in terms of trying to place hard-to-place marginal kidneys. In general, there is evidence that transplant recipients can benefit from receiving a lower quality marginal kidney, because it still reduces their time on dialysis and therefore improves their quality of life [30], [31].

In a qualitative study on the AI onboarding needs of pathologists, users expressed an interest in the AI system providing both conservative and liberal predictions for outliers or edge cases to be consistent with the process of asking for a second opinion. In general, clinicians tend to prefer second opinions from doctors with a similar risk posture [37].

3.4. *Customizing AI predictions for specific cases*

On top of personalization, there could be added benefit in customizing the information from an AI depending on the case. This tailoring can be accomplished (a) by predicting case difficulty or (b) through user control of the explainability information.

3.4.1. *Case difficulty*

In the stakeholder engagement, transplant surgeons perceived the AI as most useful for difficult decisions, where they would want a second opinion. For simple cases, the AI provided little added benefit above their existing expertise. For example, transplant surgeons suggested that the AI could determine how much explainability information to provide based on donor characteristics. For a young healthy donor, they did not need AI support to decide to transplant. In contrast, OPOs perceived the explainability information as more generally useful and wanted to always review it. OPOs have a more process-oriented task requiring justification to switch to an accelerated process, and that justification could be provided by XAI.

In some cases, the AI may be able to predict whether a case is difficult, for example if it is an outlier or edge case. In these situations, users need additional explainability information because the AI is also more likely to be incorrect and therefore is less trustworthy. Experts are often able to discern when it is a difficult decision. For example, users' perceived trust and perceived performance of AI decreased as the task difficulty increased [111].

In the context of autonomous driving, users often compared the vehicle's behavior to their own, mentioning how they would have performed the task differently, especially in cases where their competence was higher than the vehicle [49]. The appropriate amount of explainability may also depend on the specific task and therefore the role of the human user [46]. In a pediatric ICU context, physicians preferred detailed information for data exploration while nurses preferred more precise, actionable information [48].

3.4.2. Control of explainability information

Alternatively, if users can identify a difficult case, they may be able to control the use of explainability information. In this situation, rather than identifying when the AI is incorrect, the user is focused on identifying when the AI is likely to be less useful. In the stakeholder engagement, transplant professionals and OPO coordinators wanted to control the level of explainability information provided based on the complexity of each case. They preferred to choose whether to view explainability information via buttons or expandable dropdowns.

Users may prefer to only view explainability information if requested [66]. There is some evidence that this is an effective strategy. In reviewing medical images, lay users were better able to recognize correct and incorrect AI recommendations when explainability information was provided separately rather than on top of the image [112]. Similarly, pediatric ICU clinicians also expressed a preference to only access some explainability features when needed, whereas they wanted others to be provided by default [48]. In another study, medical experts mentioned that having AI predictions for a typical use case along with the current case will improve their decision making [13]. For pathologists reviewing images, trust in the system increased when they could guide the system towards the right direction by selecting similar instances [113]. Similarly, users preferred a system they can modify and control even if only a small number of modifications are allowed [114].

However, there is a risk of information overload. Users supervising an unmanned aerial vehicle delivering packages reported that basic textual explanations improved their understanding whereas users who received a fully detailed explanation were overwhelmed [115]. Users perform significantly better, take less time, and have higher confidence with simpler decision table explanations than more complex explanations [113], [114], [116], [117]. Additionally, users performed significantly better when explanations were easily accessible compared to either no explanations or more difficult to access explanations that required multiple clicks [118]. Another qualitative study found that users may benefit if XAI is contextualized based on case severity, risk posture, and time sensitivity [119]. This suggests that enabling users to customize an XAI interface may be beneficial for user adoption [120], improving task performance [118] and reducing information overload.

4. Discussion and areas for future research

This study contextualized findings from a stakeholder engagement focused on the kidney transplant placement process with the XAI literature to identify human-centered insights for XAI interaction.

4.1. Integration of AI in stakeholder decision-making

There is value in soliciting stakeholder input as early in the design process as possible to identify the appropriate role of AI in a human-AI team, evaluate heterogeneity in the human's decision-making model and understand both human and AI constraints. For the transplant case, we anticipate framing an AI as a screening tool for OPOs and as a second opinion for transplant centers, while leveraging highlighting capabilities for both use cases. In many cases, the needs of various stakeholders may be in conflict with one another and necessitate different interfaces. For example, patient perspectives are not discussed in this analysis because they have different goals and are often seeking more comprehensive information about the transplant process in general, rather than support for a single decision. In this specific case, additional stakeholder engagement is needed to determine how and when to give patients AI support for evaluating specific kidney offers. However, the patient perspective is still needed here to understand their preferences for how their doctors and other staff interact with AI tools. Future research should explore how an AI can determine recommendations based on classification thresholds in multi-stakeholder medical settings, considering whether a fixed threshold, clinicians' interpretation of AI predictions, as well as patient preferences and risk tolerances should influence decision-making [121], [122]. Additional research is also needed to understand how framing an AI as a certain role within a human-AI team influences perceptions, adoption, and performance. We expect that providing clear framing for how to interact with an AI for a particular task will improve performance by supporting a cooperative, rather than competitive, interaction. In addition, clear framing on the role of the AI may alleviate concerns about AI replacing workers and liability for negative outcomes.

Designers of AI decision support systems need to determine whether users are most likely to benefit from a system that mirrors their decision-making process or not. AI predictions may be more helpful when they are consistent with the human's existing decision-making process. This consistency may improve users' ability to evaluate the quality of the AI prediction. Other decision support tools in the transplant space have focused on predicting metrics that inform the final accept or decline decision, such as the time to better offer, probability of graft survival, and patient mortality [123], [124]. In future research, it would be valuable to compare providing decision support by using these types of metrics versus directly predicting the final decision.

4.2. Provision of system-level and prediction-level information

To date, there is significantly more research being conducted on how to communicate prediction-level explainability information, rather than system-level explainability information. In many cases, system-level information is framed as a disclosure, such as Google Model Cards, IBM AI FactSheets, and the Dataset Nutrition Label [125–128]. Literature suggests that both system-level and prediction-level information are necessary for users to build an appropriate mental model of the system. To improve these communications, AI interface designers and researchers may find it valuable to conduct mental models research to formally characterize human users' mental models of both the task at hand as well as the AI model [49]. Identifying misconceptions may be valuable for identifying the most important information to communicate [129], especially in healthcare where AI's relative newness leads to misconceptions [130] and lack of basic AI knowledge among physicians [131]. Research is needed to determine the importance of system-level information in helping users appropriately trust AI models and support global model reasoning. For example, system-level information about how the model was trained could help users identify outliers or edge cases where the model may not be as good at making predictions. System-level methods, such as feature importance and decision trees, tend to improve users' trust and performance [91], [116], [132], [133] and could also help with adoption of the system [134]. Future research is also needed to determine the most effective method and how often to provide system-level XAI information. Objectively evaluating the user's understanding of the AI system's performance metrics, training process, and boundary conditions could help users build an appropriate mental model and trust in the system. Future research is required to determine how frequently users need to review system-level information to maintain appropriate trust and an accurate mental model of the system and its limitations.

In addition to system-level information, users may benefit from prediction-level information. Uncertainty metrics may help users understand when to trust AI predictions, as high uncertainty suggests that the AI may be less trustworthy for a particular case [135]. For explainability, a wide range of methods have been tested across textual, numerical, and visual formats. Feature relevance may be a good starting point to improve task performance and user trust. Human-AI teams rarely achieve complimentary performance. This is due in part to other challenges previously mentioned, such as poor mental models and inadequate onboarding. Studies suggest that XAI provided in a combination of textual, numerical, and visual formats improve users' task performance, especially for novice users [136], [137]. Expert users may benefit more from having control of what XAI information they see. Future research should focus on developing a multi-step empirical process of initially educating the stakeholders about the AI with system-level information and then investigating the effects of various prediction-level XAI information on task performance and user trust. Showing system-level information at regular intervals or with prediction-level information may help improve the human-AI team's task performance.

4.3. Interaction of user expertise and level of XAI information

For personalization, it may be valuable to adapt an AI operation and/or interface to better account for the user's expertise level and/or decision-making threshold. Novice stakeholders may benefit from more explainability information in multiple modes (e.g., text and visual), while experts may benefit from more interactive interfaces to support engagement. Most research has focused on novice users because they are more accessible. It can be challenging to recruit busy professionals to provide an expert perspective in many studies. In these cases, it may be valuable to identify an alternative task that can be conducted by the public with consistent characteristics to the target task. In the transplant context, we have conducted studies in analogous domains, such as basketball betting [138] and image identification [87]. While not perfect, this can support theory development and refine the design before testing with the target population.

4.4. Customization of system- and prediction-level information based on user and task

For customization, it may be valuable to tailor AI communications depending on how easy or difficult a decision is for the user. This may be achieved automatically via an indicator that a particular case is an outlier. This may prompt the user to spend extra time examining the case, improving decision-making regardless of the quality of the AI prediction [132]. Alternatively, it should be valuable to build in flexibility in an AI interface, if it does not reduce performance [139]. More research is needed to understand when and how a user can self-manage the information that they get from an AI, whether that is explainability information or opting-in to a prediction in the first place. It may be difficult to anticipate the appropriate amount of AI assistance, which may vary based on a dynamic process that is sensitive to the individual, task, and environment. Given concerns from the literature that users often prefer communications that do not improve performance, it may be valuable to design strong defaults to encourage effective use of an AI support tool. It is also possible that experts are better positioned to do this self-management than novice users. Specifically, future research should determine the level of XAI information required, for easy versus difficult use cases, and experts versus novice stakeholders, regardless of user preference.

5. Conclusion

Using stakeholder engagement to guide a literature review is an effective strategy for identifying new areas for research. Four primary themes emerged from this process related to 1) use of AI predictions, 2) information included in AI predictions, 3) personalization of AI prediction for different groups, and 4) customizing AI prediction for specific cases. One of the primary findings here is the potential value of flexibility when implementing AI decision support in the real world. More research is needed to understand when user control is appropriate, and for which users performing which tasks. While AI assistance has the potential to improve decision-making performance and efficiency, it may also burden users by providing too much information for cases where it is unhelpful or lead to over-reliance and inappropriate trust. Further research is needed to help users manage and control their interactions with AI decision support.

Funding

This work was supported by National Science Foundation (NSF) grants (#2222801 & #2026324).

CRediT authorship contribution statement

Harishankar V. Subramanian: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Casey Canfield:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review & editing. **Daniel B. Shank:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Casey Canfield reports financial support was provided by National Science Foundation.

Acknowledgments

We thank our colleagues and collaborators at Saint Louis University who have provided critical insight into the kidney transplant process, Krista Lentine, Mark Schnitzler, Henry Randall, Jason Eberl, and Michael Miller as well as those at the United Network for Organ Sharing, Stephanie Rose, Brendon Cummiskey, Samantha Noreen, and Laura Cartwright. We also thank our colleagues at Missouri S&T, including Cihan Dagli, Venkata Sriram Siddhardh Nadendla, Lirim Ashiku, Richard Threlkeld, Mukund Telukunta, and Rachel Dzieran. Also, we thank Hannah Elder and Casey Hines for their roles in designing and conducting the stakeholder interviews. We also thank the other members of the Canfield Lab for their input, Ankit Agarwal, Victoria Kraemer, Eyuel Getahun, and Elham Babaee. Lastly, we thank our stakeholders from Organ Procurement Organizations, Transplant Centers, and recipients for their input – Sage Bailey, Diane Brockmeier, Matthew Cooper, Meelie DebRoy, Kevin Doerschug, Jameson Forster, Kevin Fowler, Brittney Gabris, Patrick Gee, Christie Gooden, Amanda Grandinetti, Darla Granger, Stephen Gray, Michael Harmon, Holly Jackson, Nichole Jefferson, Martin Jendrisak, Andrea Koehler, Melissa Lichtenberger, Roslyn Mannon, Lori Markham, Gary Marklin, Mae McCandless, Marc Melcher, Whitni Noyes, Angela Pearson, Sam Pederson, Glenda Roberts, Richard Rothweiler, Brian Scheller, John Stalbaum, Meghan Stephenson, Silla Sumerlin, Curtis Warfield, Jason Wellen, David White, Harry Wilkins, and Cody Wooley.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102780. This includes additional figures and a summary of the reviewed papers. Data collection materials and aggregated results are available on Open Science Framework at https://osf.io/ju9x3/.

References

[1] C. Ross and I. Swetlitz, “IBM's Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show,” STAT+, pp. 1–10, Jul. 25, 2018.

[2] J. Angin, J. Larson, M. Surya, and L. Kirchner, “Machine Bias — ProPublica,” 2016, [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] A. Chouldechova, “Fair prediction with disparate impact: a study of bias in recidivism prediction instruments,” Big Data, vol. 5, no. 2, pp. 153–163, Jun. 2017, doi:https://doi.org/10.1089/big.2016.0047.

[4] M. Livingston, “Preventing racial bias in federal AI,” J Sci Policy Gov, vol. 16, no. 02, May 2020, doi:10.38126/JSPG160205.

[5] C. Manresa-Yee, “Advances in XAI: explanation interfaces in healthcare,” in *Handbook of artificial intelligence in healthcare*, vol. 212, C. Manresa-Yee, M. F. Roig-Maimó, S. Ramis, and R. Mas-Sansó, Eds., in Intelligent Systems Reference Library, vol. 212. , Cham: Springer International Publishing, 2022, pp. 357–369. doi:https://doi.org/10.1007/978-3-030-83620-7_15.

[6] Antoniadi AM, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci May 2021;11(11):5088. https://doi.org/10.3390/app11115088.

[7] Hassan A, Abdulhak MAA, Bin Sulaiman R, Kahtan H. User centric explanations: a breakthrough for explainable models. In: 2021 international conference on information technology (ICIT). Amman, Jordan: IEEE; Jul. 2021. p. 702–7. https://doi.org/10.1109/ICIT52682.2021.9491641.

[8] T. W. Concannon *et al.*, “A systematic review of stakeholder engagement in comparative effectiveness and patient-centered outcomes research,” J Gen Intern Med, vol. 29, no. 12, pp. 1692–1701, Dec. 2014, doi:https://doi.org/10.1007/s11606-014-2878-x.

[9] S. Hepenstal and D. McNeish, “Explainable artificial intelligence: what do you need to know?,” in *Augmented cognition. theoretical and technological approaches*, D. D. Schmorrow and C. M. Fidopiastis, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 266–275. doi:https://doi.org/10.1007/978-3-030-50353-6_20.

[10] A. K. M. Nor, S. R. Pedapati, M. Muhammad, and V. Leiva, “Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses,” Sensors, vol. 21, no. 23, p. 8020, Dec. 2021, doi:https://doi.org/10.3390/s21238020.

[11] Jagosh J, et al. Uncovering the benefits of participatory research: implications of a realist review for health research and practice. Milbank Q Jun. 2012;90(2): 311–46. https://doi.org/10.1111/j.1468-0009.2012.00665.x.

[12] B. Roehr, “More stakeholder engagement is needed to improve quality of research, say US experts,” BMJ, vol. 341, no. aug03 1, pp. c4193–c4193, Aug. 2010, doi:https://doi.org/10.1136/bmj.c4193.

[13] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. Van Den Bosch, “Human-centered XAI: developing design patterns for explanations of clinical decision support systems,” Int J Hum-Comput Stud, vol. 154, p. 102684, Oct. 2021, doi:https://doi.org/10.1016/j.ijhcs.2021.102684.

[14] Spinuzzi C. The methodology of participatory design. Tech Commun 2005;52(2): 163–74.

[15] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, “Explaining deep neural networks and beyond: a review of methods and applications,” Proc IEEE, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi:https://doi.org/10.1109/JPROC.2021.3060483.

[16] Wang X, Yin M. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In: 26th international conference on intelligent user interfaces. College Station TX USA: ACM, Apr; 2021. p. 318–28. https://doi.org/10.1145/3397481.3450650.

[17] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, “Explaining deep neural networks: a survey on the global interpretation methods,” Neurocomputing, vol. 513, pp. 165–180, Nov. 2022, doi:https://doi.org/10.1016/j.neucom.2022.09.129.

[18] M. T. Keane and E. M. Kenny, “How case-based reasoning explains neural networks: a theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems,” in *Case-based reasoning research and development*, vol. 11680, in Lecture Notes in Computer Science, vol. 11680., Cham: Springer International Publishing, 2019, pp. 155–171. doi:https://doi.org/10.1007/978-3-030-29249-2_11.

[19] D. Gunning, E. Vorm, Y. Wang, and M. Turek, “DARPA's explainable AI (XAI) program: a retrospective,” Appl AI Lett, vol. 2, no. 4, pp. 1–12, Nov. 2021, doi:https://doi.org/10.1002/ail2.61.

[20] Verhagen RS, Neerincx MA, Tielman ML. A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. In: Calvaresi D, Najjar A, Winikoff M, Framling K, editors. Explainable and transparent AI and multi-agent systems. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2021. p. 119–38. https://doi.org/10.1007/978-3-030-82017-6_8.

[21] R. O. Alabi *et al.*, “Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future—a systematic review,” Artif Intell Med, vol. 115, p. 102060, May 2021, doi:https://doi.org/10.1016/j.artmed.2021.102060.

[22] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, “Evaluating the quality of machine learning explanations: a survey on methods and metrics,” Electronics, vol. 10, no. 5, p. 593, Mar. 2021, doi:https://doi.org/10.3390/electronics10050593.

[23] Z. Zhang, D. Citardi, D. Wang, Y. Genc, J. Shan, and X. Fan, “Patients' perceptions of using artificial intelligence (AI)-based technology to comprehend radiology imaging data,” *Health Informatics J.*, vol. 27, no. 2, Apr. 2021, doi:https://doi.org/10.1177/14604582211011215.

[24] Amershi S, et al. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. Glasgow Scotland Uk: ACM; May 2019. p. 1–13. https://doi.org/10.1145/3290605.3300233.

[25] L. Gates and D. Leake, “Evaluating CBR explanation capabilities: survey and next steps,” in *CEUR Workshop Proceedings*, 2021, pp. 40–51.

[26] Bhatt U, et al. Uncertainty as a form of transparency: measuring, communicating, and using uncertainty. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society. Virtual Event USA: ACM; Jul. 2021. p. 401–13. https://doi.org/10.1145/3461702.3462571.

[27] D. Wang, W. Zhang, and B. Y. Lim, “Show or suppress? Managing input uncertainty in machine learning model explanations,” Artif Intell, vol. 294, p. 103456, May 2021, doi:https://doi.org/10.1016/j.artint.2021.103456.

[28] “UNOS data and transplant statistics: Organ Donation Data.” [Online]. Available: https://unos.org/data/.

[29] D. A. Axelrod *et al.*, “An economic assessment of contemporary kidney transplant practice,” Am J Transplant, vol. 18, no. 5, pp. 1168–1176, May 2018, doi:https://doi.org/10.1111/ajt.14702.

[30] C. L. Jay, K. Washburn, P. G. Dean, R. A. Helmick, J. A. Pugh, and M. D. Stegall, “Survival benefit in older patients associated with earlier transplant with high KDPI kidneys,” Transplantation, vol. 101, no. 4, pp. 867–872, Apr. 2017, doi:https://doi.org/10.1097/TP.0000000000001405.

[31] A. B. Massie, X. Luo, E. K. H. Chow, J. L. Alejo, N. M. Desai, and D. L. Segev, “Survival benefit of primary deceased donor transplantation with high-KDPI kidneys,” Am J Transplant, vol. 14, no. 10, pp. 2310–2316, Oct. 2014, doi:https://doi.org/10.1111/ajt.12830.

[32] O. Aubert *et al.*, “Disparities in acceptance of deceased donor kidneys between the United States and France and estimated effects of increased US acceptance,” JAMA Intern Med, vol. 179, no. 10, p. 1365, Oct. 2019, doi:https://doi.org/10.1001/jamainternmed.2019.2322.

[33] S. Mohan *et al.*, “Factors leading to the discard of deceased donor kidneys in the United States,” Kidney Int, vol. 94, no. 1, pp. 187–198, Jul. 2018, doi:https://doi.org/10.1016/j.kint.2018.02.016.

[34] J. R. F. Narvaez, J. Nie, K. Noyes, M. Leeman, and L. K. Kayler, “Hard-to-place kidney offers: donor- and system-level predictors of discard,” Am J Transplant, vol. 18, no. 11, pp. 2708–2718, Nov. 2018, doi:https://doi.org/10.1111/ajt.14712.

[35] Cargo M, Mercer SL. The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 2008;29(1):25–50.

[36] J. Harris, L. Croot, J. Thompson, and J. Springett, “How stakeholder participation can contribute to systematic reviews of complex interventions,” J Epidemiol Community Health, vol. 70, no. 2, pp. 207–214, Feb. 2016, doi:https://doi.org/10.1136/jech-2015-205701.

[37] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, “‘Hello AI’: uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–24, Nov. 2019, doi:https://doi.org/10.1145/3359206.

[38] R. Threlkeld *et al.*, “Reducing kidney discard with artificial intelligence decision support: the need for a transdisciplinary systems approach,” Curr Transplant Rep, vol. 8, no. 4, pp. 263–271, Dec. 2021, doi:https://doi.org/10.1007/s40472-021-00351-0.

[39] L. Bowker and D. Fisher, “Computer-aided translation,” in *Handbook of translation studies*, Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2010, pp. 60–65.

[40] D. R. Thomas, “A general inductive approach for analyzing qualitative evaluation data,” Am J Eval, vol. 27, no. 2, pp. 237–246, Jun. 2006, doi:https://doi.org/10.1177/1098214005283748.

[41] H. J. Schielke, J. L. Fishman, K. Osatuke, and W. B. Stiles, “Creative consensus on interpretations of qualitative data: the Ward method,” Psychother Res, vol. 19, no. 4–5, pp. 558–565, Jul. 2009, doi:https://doi.org/10.1080/10503300802621180.

[42] G. Baudry, F. Delrue, J. Legrand, J. Pruvost, and T. Vallée, “The challenge of measuring biofuel sustainability: a stakeholder-driven approach applied to the French case,” Renew Sust Energ Rev, vol. 69, pp. 933–947, Mar. 2017, doi:https://doi.org/10.1016/j.rser.2016.11.022.

[43] A. C. Tricco *et al.*, “PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation,” Ann Intern Med, vol. 169, no. 7, pp. 467–473, Oct. 2018, doi:https://doi.org/10.7326/M18-0850.

[44] Z. Munn *et al.*, “What are scoping reviews? Providing a formal definition of scoping reviews as a type of evidence synthesis,” JBI Evid Synth, vol. 20, no. 4, pp. 950–952, Apr. 2022, doi:10.11124/JBIES-21-00483.

[45] M. Dragoni, I. Donadello, and C. Eccher, “Explainable AI meets persuasiveness: translating reasoning results into behavioral change advice,” Artif Intell Med, vol. 105, p. 101840, May 2020, doi:https://doi.org/10.1016/j.artmed.2020.101840.

[46] A. Richardson and A. Rosenfeld, “A survey of interpretability and explainability in human-agent systems,” in *XAI Workshop on Explainable* Artif Intell, Jul. 2018, pp. 137–143.

[47] A. D. Jeffery, L. L. Novak, B. Kennedy, M. S. Dietrich, and L. C. Mion, "Participatory design of probability-based decision support tools for in-hospital nurses," J Am Med Inform Assoc, vol. 24, no. 6, pp. 1102–1110, Nov. 2017, doi: https://doi.org/10.1093/jamia/ocx060.

[48] A. J. Barda, C. M. Horvat, and H. Hochheiser, "A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare," BMC Med Inform Decis Mak, vol. 20, no. 1, p. 257, Dec. 2020, doi:https://doi.org/10.1186/s12911-020-01276-x.

[49] Wiegand G, Eiband M, Haubelt M, Hussmann H. 'I'd like an explanation for that!'Exploring reactions to unexpected autonomous driving. In: 22nd international conference on human-computer interaction with Mobile devices and services. Oldenburg Germany: ACM; Oct. 2020. p. 1–11. https://doi.org/10.1145/3379503.3403554.

[50] Omeiza D, Kollnig K, Web H, Jirotka M, Kunze L. Why not explain? Effects of explanations on human perceptions of autonomous driving. In: 2021 IEEE international conference on advanced robotics and its social impacts (ARSO). Tokoname, Japan: IEEE; Jul. 2021. p. 194–9. https://doi.org/10.1109/ARSO51874.2021.9542835.

[51] Chromik M, Schuessler M. A taxonomy for human subject evaluation of black-box explanations in XAI. Proceedings of ExSS-ATEC 2020.

[52] Bansal G, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Yokohama Japan: ACM; May 2021. p. 1–16. https://doi.org/10.1145/3411764.3445717.

[53] Zhang Y, Liao QV, Bellamy RKE. 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. Barcelona Spain: ACM; Jan. 2020. p. 295–305. https://doi.org/10.1145/3351095.3372852.

[54] M. Barah, V. Kilambi, J. J. Friedewald, and S. Mehrotra, "Implications of accumulated cold time for US kidney transplantation offer acceptance," Clin J Am Soc Nephrol, vol. 17, no. 9, pp. 1353–1362, Sep. 2022, doi:https://doi.org/10.2215/CJN.01600222.

[55] B. L. Kasiske *et al.*, "The role of procurement biopsies in acceptance decisions for kidneys retrieved for transplant," Clin J Am Soc Nephrol, vol. 9, no. 3, pp. 562–571, Mar. 2014, doi:https://doi.org/10.2215/CJN.07610713.

[56] K. L. Lentine, B. Kasiske, and D. A. Axelrod, "Procurement biopsies in kidney transplantation: more information may not lead to better decisions," J Am Soc Nephrol JASN, vol. 32, no. 8, pp. 1835–1837, Aug. 2021, doi:https://doi.org/10.1681/ASN.2021030403.

[57] Threlkeld R, et al. AI-enabled digital support to increase placement of hard-to-place deceased donor kidneys. Am J Transplant 2023;23(6):S815–6.

[58] J. Hwang, T. Lee, H. Lee, and S. Byun, "A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study," J Med Internet Res, vol. 24, no. 1, p. e28659, Jan. 2022, doi:https://doi.org/10.2196/28659.

[59] L. Sanneman and J. A. Shah, "A situation awareness-based framework for design and evaluation of explainable AI," in *Explainable, transparent autonomous agents and multi-agent systems*, vol. 12175, in Lecture Notes in Computer Science, vol. 12175. , Cham: Springer International Publishing, 2020, pp. 94–110. doi:https://doi.org/10.1007/978-3-030-51924-7_6.

[60] R. Larasati, A. De Liddo, and E. Motta, "AI healthcare system interface: explanation design for non-expert user trust," in *ACMIUI-WS 2021: Joint Proceedings of the ACM IUI 2021 Workshops*, D. Glowacka and V. Krishnamurthy, Eds., CEUR Workshop Proceedings, Apr. 2021.

[61] Das D, Chernova S. Leveraging rationales to improve human task performance. In: *Proceedings of the 25th international conference on intelligent user interfaces*, in IUI '20. New York, NY, USA: Association for Computing Machinery; Mar. 2020. p. 510–8. https://doi.org/10.1145/3377325.3377512.

[62] M. Ribera and A. Lapedriza, "Can we do better explanations? A proposal of user-centered explainable AI," in *CEUR Workshop Proceedings*, 2019, p. 7.

[63] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," Inf Fusion, vol. 76, pp. 89–106, Dec. 2021, doi:https://doi.org/10.1016/j.inffus.2021.05.009.

[64] Cheng H-F, et al. Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In: Proceedings of the 2019 CHI conference on human factors in computing systems. Glasgow Scotland Uk: ACM; May 2019. p. 1–12. https://doi.org/10.1145/3290605.3300789.

[65] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, "Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies," Artif Intell, vol. 294, p. 103459, May 2021, doi: https://doi.org/10.1016/j.artint.2021.103459.

[66] M. Chromik and A. Butz, "Human-XAI interaction: a review and design principles for explanation user interfaces," in *Human-computer interaction – INTERACT* 2021, Vol. 12933, in lecture notes in computer science, vol. 12933. , Cham: Springer International Publishing, 2021, pp. 619–640. doi:https://doi.org/10.1007/978-3-030-85616-8_36.

[67] Jesus S, et al. How can I choose an explainer?: an application-grounded evaluation of post-hoc explanations. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Virtual Event Canada: ACM; Mar. 2021. p. 805–15. https://doi.org/10.1145/3442188.3445941.

[68] V. L. Pop, A. Shrewsbury, and F. T. Durso, "Individual differences in the calibration of trust in automation," Hum Factors J Hum Factors Ergon Soc, vol. 57, no. 4, pp. 545–556, Jun. 2015, doi:https://doi.org/10.1177/0018720814564422.

[69] L. Arbelaez Ossa, M. Rost, G. Lorenzini, D. M. Shaw, and B. S. Elger, "A smarter perspective: learning with and from AI-cases," Artif Intell Med, vol. 135, p. 102458, Jan. 2023, doi:https://doi.org/10.1016/j.artmed.2022.102458.

[70] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," IEEE Access, vol. 9, pp. 153316–153348, 2021, doi:https://doi.org/10.1109/ACCESS.2021.3127881.

[71] Beauxis-Aussalet E, Van Doorn J, Hardman L. Supporting end-user understanding of classification errors. In: Proceedings of the 36th European conference on cognitive ergonomics. Utrecht Netherlands: ACM; Sep. 2018. p. 1–8. https://doi.org/10.1145/3232078.3232096.

[72] H. Shen, H. Jin, Á. A. Cabrera, A. Perer, H. Zhu, and J. I. Hong, "Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance," Proc ACM Hum-Comput Interact, vol. 4, no. CSCW2, pp. 1–22, Oct. 2020, doi:https://doi.org/10.1145/3415224.

[73] Park S, et al. Impact of expectation and performance on the user experience of AI systems. ICIC International 学会 2022. https://doi.org/10.24507/icicelb.13.01.73.

[74] K. Yu, S. Berkovsky, D. Conway, R. Taib, J. Zhou, and F. Chen, "Do I trust a machine? Differences in user trust based on system performance," in *Human and machine learning: visible, explainable, trustworthy and transparent*, J. Zhou and F. Chen, Eds., in Human–computer interaction series., Cham: Springer International Publishing, 2018, pp. 245–264. doi:https://doi.org/10.1007/978-3-319-90403-0_12.

[75] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust." arXiv, Jul. 26, 2019. [Online]. Available: http://arxiv.org/abs/1907.12652.

[76] 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. Glasgow Scotland Uk: ACM; May 2019. p. 1–12. https://doi.org/10.1145/3290605.3300509.

[77] Kizilcec RF. How much information?: effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI conference on human factors in computing systems. San Jose California USA: ACM; May 2016. p. 2390–5. https://doi.org/10.1145/2858036.2858402.

[78] J. Zhou, H. Hu, Z. Li, K. Yu, and F. Chen, "Physiological indicators for user trust in machine learning with influence enhanced fact-checking," in *Machine learning and knowledge extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 94–113. doi:https://doi.org/10.1007/978-3-030-29726-8_7.

[79] Suresh H, Lao N, Liccardi I. Misplaced trust: measuring the interference of machine learning in human decision-making. In: 12th ACM conference on Web science. Southampton United Kingdom: ACM; Jul. 2020. p. 315–24. https://doi.org/10.1145/3394231.3397922.

[80] L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," Requir Eng, vol. 25, no. 4, pp. 493–514, Dec. 2020, doi:https://doi.org/10.1007/s00766-020-00333-1.

[81] C.-H. Tsai and P. Brusilovsky, "The effects of controllability and explainability in a social recommender system," User Model User-Adapt Interact, vol. 31, no. 3, pp. 591–627, Jul. 2021, doi:https://doi.org/10.1007/s11257-020-09281-5.

[82] Tsai C-H, You Y, Gui X, Kou Y, Carroll JM. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Yokohama Japan: ACM; May 2021. p. 1–17. https://doi.org/10.1145/3411764.3445101.

[83] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, "'Let me explain!': exploring the potential of virtual agents in explainable AI interaction design," J Multimodal User Interfaces, vol. 15, no. 2, pp. 87–98, Jun. 2021, doi:https://doi.org/10.1007/s12193-020-00332-0.

[84] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: beware of inmates running the asylum or: how I learnt to stop worrying and love the social and behavioural sciences." arXiv, Dec. 04, 2017. [Online]. Available: http://arxiv.org/abs/1712.00547.

[85] F. M. Calisto, C. Santiago, N. Nunes, and J. C. Nascimento, "BreastScreening-AI: evaluating medical intelligent agents for human-AI interactions," Artif Intell Med, vol. 127, p. 102285, May 2022, doi:https://doi.org/10.1016/j.artmed.2022.102285.

[86] Alqaraawi A, Schuessler M, Weiß P, Costanza E, Berthouze N. Evaluating saliency map explanations for convolutional neural networks: a user study. In: Proceedings of the 25th international conference on intelligent user interfaces. Cagliari Italy: ACM; Mar. 2020. p. 275–85. https://doi.org/10.1145/3377325.3377519.

[87] Khurana A, Alamzadeh P, Chilana PK. ChatrEx: designing explainable Chatbot interfaces for enhancing usefulness, transparency, and trust. In: *2021 IEEE symposium on visual languages and human-centric computing (VL/HCC)*, St Louis. MO, USA: IEEE; Oct. 2021. p. 1–11. https://doi.org/10.1109/VL/HCC51201.2021.9576440.

[88] A. V. González, G. Bansal, A. Fan, Y. Mehdad, R. Jia, and S. Iyer, "Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, 2021, pp. 1103–1116. doi:10.18653/v1/2021.findings-acl.95.

[89] P. Tschandl *et al.*, "Human–computer collaboration for skin cancer recognition," Nat Med, vol. 26, no. 8, pp. 1229–1234, Aug. 2020, doi:https://doi.org/10.1038/s41591-020-0942-0.

[90] H. V. Subramanian, C. I. Canfield, D. B. Shank, L. Andrews, and C. H. Dagli, "Communicating uncertain information from deep learning models in human

machine teams," in *American Society for Engineering Management International Annual Conference*, American Society for Engineering Management (ASEM), 2020.

[91] Yang F, Huang Z, Scholtz J, Arendt DL. How do visual explanations foster end users' appropriate trust in machine learning?. In: Proceedings of the 25th international conference on intelligent user interfaces. Cagliari Italy: ACM; Mar. 2020. p. 189–201. https://doi.org/10.1145/3377325.3377480.

[92] Buçinca Z, Lin P, Gajos KZ, Glassman EL. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In: Proceedings of the 25th international conference on intelligent user interfaces. Cagliari Italy: ACM; Mar. 2020. p. 454–64. https://doi.org/10.1145/3377325.3377498.

[93] A. R. Akula *et al.*, "CX-ToM: counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models," iScience, vol. 25, no. 1, p. 103581, Jan. 2022, doi:https://doi.org/10.1016/j.isci.2021.103581.

[94] Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency. Atlanta GA USA: ACM; Jan. 2019. p. 29–38. https://doi.org/10.1145/3287560.3287590.

[95] Lai V, Liu H, Tan C. 'Why is "Chicago" deceptive?' Towards building model-driven tutorials for humans. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*, in CHI '20. New York, NY, USA: Association for Computing Machinery; Apr. 2020. p. 1–13. https://doi.org/10.1145/3313831.3376873.

[96] J. Wanner, "Do you really want to know why? Effects of AI-based DSS advice on human decisions," in *27th Annual Americas Conference on Information Systems,* AMCIS 2021, 2021, p. 10.

[97] H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making," Proc ACM Hum-Comput Interact, vol. 5, no. CSCW2, pp. 1–45, Oct. 2021, doi:https://doi.org/10.1145/3479552.

[98] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection," Transl Psychiatry, vol. 11, no. 1, p. 108, Feb. 2021, doi:https://doi.org/10.1038/s41398-021-01224-x.

[99] M. Merry, P. Riddle, and J. Warren, "A mental models approach for defining explainable artificial intelligence," BMC Med Inform Decis Mak, vol. 21, no. 1, p. 344, Dec. 2021, doi:https://doi.org/10.1186/s12911-021-01703-7.

[100] Chromik M, Eiband M, Buchner F, Krüger A, Butz A. I think I get your point, AI! The illusion of explanatory depth in explainable AI. In: 26th international conference on intelligent user interfaces. College Station TX USA: ACM; Apr. 2021. p. 307–17. https://doi.org/10.1145/3397481.3450644.

[101] R. Larasati, A. D. Liddo, and E. Motta, "The effect of explanation styles on user's trust," in 2020 Workshop on explainable smart systems for algorithmic transparency in emerging technologies, Mar 2020.

[102] Szymanski M, Millecamp M, Verbert K. Visual, textual or hybrid: the effect of user expertise on different explanations. In: 26th international conference on intelligent user interfaces. College Station TX USA: ACM, Apr; 2021. p. 109–19. https://doi.org/10.1145/3397481.3450662.

[103] L. K. Branting *et al.*, "Scalable and explainable legal prediction," Artif Intell Law, vol. 29, no. 2, pp. 213–238, Jun. 2021, doi:https://doi.org/10.1007/s10506-020-09273-1.

[104] T. Schrills and T. Franke, "Color for characters - effects of visual explanations of AI on trust and observability," in *Artificial Intelligence in HCI*, vol. 12217, H. Degen and L. Reinerman-Jones, Eds., in Lecture Notes in Computer Science, vol. 12217. , Cham: Springer International Publishing, 2020, pp. 121–135. doi:https://doi.org/10.1007/978-3-030-50334-5_8.

[105] C. Woodcock, B. Mittelstadt, D. Busbridge, and G. Blank, "The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: experimental study," J Med Internet Res, vol. 23, no. 11, p. e29386, Nov. 2021, doi:https://doi.org/10.2196/29386.

[106] Y. Xie, M. Chen, D. Kao, G. Gao, and X. "Anthony" Chen, "CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–13. doi:https://doi.org/10.1145/3313831.3376807.

[107] S. Hepenstal, L. Zhang, N. Kodagoda, and B. L. W. Wong, "A granular computing approach to provide transparency of intelligent systems for criminal investigations," in *Interpretable Artificial Intelligence: A Perspective of Granular Computing*, vol. 937, W. Pedrycz and S.-M. Chen, Eds., in Studies in Computational Intelligence, vol. 937. , Cham: Springer International Publishing, 2021, pp. 333–367. doi:https://doi.org/10.1007/978-3-030-64949-4_11.

[108] Levy A, Agrawal M, Satyanarayan A, Sontag D. Assessing the impact of automated suggestions on decision making: domain experts mediate model errors but take less initiative. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Yokohama Japan: ACM; May 2021. p. 1–13. https://doi.org/10.1145/3411764.3445522.

[109] Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting interpretability: Understanding data Scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Honolulu HI USA: ACM; Apr. 2020. p. 1–14. https://doi.org/10.1145/3313831.3376219.

[110] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation." arXiv, Feb. 02, 2018. [Online]. Available: http://arxiv.org/abs/1802.00682.

[111] O. Vl. Bitkina, H. Jeong, B. C. Lee, J. Park, J. Park, and H. K. Kim, "Perceived trust in artificial intelligence technologies: a preliminary study," Hum Factors Ergon Manuf Serv Ind, vol. 30, no. 4, pp. 282–290, Jul. 2020, doi:https://doi.org/10.1002/hfm.20839.

[112] S. Knapič, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain," Mach Learn Knowl Extr, vol. 3, no. 3, pp. 740–770, Sep. 2021, doi:https://doi.org/10.3390/make3030037.

[113] Cai CJ, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI conference on human factors in computing systems. Glasgow Scotland Uk: ACM; May 2019. p. 1–14. https://doi.org/10.1145/3290605.3300234.

[114] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them," Manag Sci, vol. 64, no. 3, pp. 1155–1170, Mar. 2018, doi:https://doi.org/10.1287/mnsc.2016.2643.

[115] Mualla Y, Tchappi I, Najjar A, Kampik T, Galland S, Nicolle C. Human-agent explainability: an experimental case study on the filtering of explanations. In: Proceedings of the 12th international conference on agents and artificial intelligence. Valletta, Malta: SCITEPRESS - Science and Technology Publications; 2020. p. 378–85. https://doi.org/10.5220/0009382903780385.

[116] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," Decis Support Syst, vol. 51, no. 1, pp. 141–154, Apr. 2011, doi:https://doi.org/10.1016/j.dss.2010.12.003.

[117] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: the role of mental models in human-AI team performance," Proc AAAI Conf Hum Comput Crowdsourcing, vol. 7, pp. 2–11, Oct. 2019, doi:https://doi.org/10.1609/hcomp.v7i1.5285.

[118] Bigras E, et al. In AI we trust: characteristics influencing assortment planners' perceptions of AI based recommendation agents. In: Nah FF-H, Xiao BS, editors. HCI in business, government, and organizations. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2018. p. 3–16. https://doi.org/10.1007/978-3-319-91716-0_1.

[119] Xie Y, Chen A, Gao G. Outlining the design space of explainable intelligent systems for medical diagnosis. ArXiv Prepr 2019;ArXiv190206019.

[120] U. Bhatt, M. Andrus, A. Weller, and A. Xiang, "Machine learning explainability for external stakeholders." arXiv, Jul. 10, 2020. [Online]. Available: http://arxiv.org/abs/2007.05408.

[121] J. Birch, K. A. Creel, A. K. Jha, and A. Plutynski, "Clinical decisions using AI must consider patient values," Nat Med, vol. 28, no. 2, pp. 229–232, Feb. 2022, doi:doi: https://doi.org/10.1038/s41591-021-01624-y.

[122] C. Barata *et al.*, "A reinforcement learning model for AI-based decision support in skin cancer," Nat Med, vol. 29, no. 8, Art. no. 8, Aug. 2023, doi:https://doi.org/10.1038/s41591-023-02475-5.

[123] J. D. Schold, A. M. Huml, E. D. Poggio, P. P. Reese, and S. Mohan, "A tool for decision-making in kidney transplant candidates with poor prognosis to receive deceased donor transplantation in the United States," Kidney Int, vol. 102, no. 3, pp. 640–651, Sep. 2022, doi:https://doi.org/10.1016/j.kint.2022.05.025.

[124] A. Wey *et al.*, "A kidney offer acceptance decision tool to inform the decision to accept an offer or wait for a better kidney," Am J Transplant, vol. 18, no. 4, pp. 897–906, Apr. 2018, doi:https://doi.org/10.1111/ajt.14506.

[125] M. Arnold *et al.*, "FactSheets: increasing trust in AI services through supplier's declarations of conformity." arXiv, Feb. 07, 2019. [Online]. Available: http://arxiv.org/abs/1808.07261.

[126] Chmielinski KS, et al. The dataset nutrition label (2nd gen): leveraging context to mitigate harms in artificial intelligence. arXiv, Mar 2022;10 [Online]. Available: http://arxiv.org/abs/2201.03954.

[127] Mitchell M, et al. Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. Atlanta GA USA: ACM; Jan. 2019. p. 220–9. https://doi.org/10.1145/3287560.3287596.

[128] S. Sabhlok, "Seamlessly govern AI models with AI factsheets and IBM OpenPages | by Shashank Sabhlok | IBM data science in practice | medium," IBM Data Science in Practice.

[129] B. B. Johnson, "Risk communication: a mental models approach," Risk Anal, vol. 22, no. 4, pp. 813–814, Aug. 2002, doi:https://doi.org/10.1111/0272-4332.00071.

[130] I. A. Scott, S. M. Carter, and E. Coiera, "Exploring stakeholder attitudes towards AI in clinical practice," BMJ Health Care Inform, vol. 28, no. 1, p. e100450, Dec. 2021, doi:https://doi.org/10.1136/bmjhci-2021-100450.

[131] Chen M, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. Front Med 2022;9. https://doi.org/10.3389/fmed.2022.990604.

[132] Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Yokohama Japan: ACM; May 2021. p. 1–52. https://doi.org/10.1145/3411764.3445315.

[133] Z. Zhang, Y. Genc, D. Wang, M. E. Ahsen, and X. Fan, "Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems," J Med Syst, vol. 45, no. 6, p. 64, Jun. 2021, doi:https://doi.org/10.1007/s10916-021-01743-6.

[134] S. V. Kovalchuk, G. D. Kopanitsa, I. V. Derevitskii, G. A. Matveev, and D. A. Savitskaya, "Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability," J Biomed Inform, vol. 127, p. 104013, Mar. 2022, doi:https://doi.org/10.1016/j.jbi.2022.104013.

[135] T. J. Loftus *et al.*, "Uncertainty-aware deep learning in healthcare: a scoping review," PLOS Digit Health, vol. 1, no. 8, p. e0000085, Aug. 2022, doi:https://doi.org/10.1371/journal.pdig.0000085.

[136] H. V. Subramanian, C. Canfield, D. B. Shank, and M. Kinnison, "Combining uncertainty information with AI recommendations supports calibration with domain knowledge," J Risk Res, vol. 26, no. 10, pp. 1137–1152, Oct. 2023, doi: https://doi.org/10.1080/13669877.2023.2259406.

[137] Herm L-V. Impact of explainable AI on cognitive load: insights from an empirical study. arXiv, Apr 2023;18. https://doi.org/10.48550/arXiv.2304.08861.

[138] H. Elder, C. Canfield, D. B. Shank, T. Rieger, and C. Hines, "Knowing when to pass: the effect of AI reliability in risky decision contexts," Hum Factors J Hum Factors Ergon Soc, p. 001872082211006, May 2022, doi:https://doi.org/10.1177/00187208221100691.

[139] L. Rundo, R. Pirrone, S. Vitabile, E. Sala, and O. Gambino, "Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine," J Biomed Inform, vol. 108, p. 103479, Aug. 2020, doi:https://doi.org/10.1016/j.jbi.2020.103479.