Sensitivity to Risk Profiles of Users When Developing AI Systems

Robin Cohen, Rishav Raj Agarwal(✉), Dhruv Kumar, Alexandre Parmentier, and Tsz Him Leung

School of Computer Science, University of Waterloo, Waterloo, Canada
{rcohen,rragarwal,d35kumar,aparment,th4heung}@uwaterloo.ca

Abstract. The AI community today has renewed concern about the social implications of the models they design, imagining the impact of deployed systems. One thrust has been to reflect on issues of fairness and explainability before the design process begins. There is increasing awareness as well of the need to engender trust from users, examining the origins of mistrust as well as the value of multiagent trust modelling solutions. In this paper, we argue that social AI efforts to date often imagine a homogenous user base and those models which do support differing solutions for users with different profiles have not yet examined one important consideration upon which trusted AI may depend: the risk profile of the user. We suggest how user risk attitudes can be integrated into approaches that try to reason about such dilemmas as sacrificing optimality for the sake of explainability. In the end, we reveal that it is challenging to be satisfying the myriad needs of users in their desire to be more comfortable accepting AI solutions and conclude that tradeoffs need to be examined and balanced. We advocate reasoning about these tradeoffs concerning user models and risk profiles, as we design the decision making algorithms of our systems.

Keywords: Position paper · Trusted AI · Risk profiles · Explainability

1 Introduction and Background

This position paper argues that deciding how to design AI systems which will be accepted by human users is a process where the specific profile of the user needs to be considered. We frame this discussion in terms of engendering trust in AI and advocate for reasoning about user risk profiles, when deciding whether to sacrifice accuracy for explainability. We note that several researchers have already suggested that differing user profiles may come into play, when studying mistrust in AI. For example, Salem et al. [20] assessed trust in robots as the participants' willingness to cooperate with the robot when it makes several unusual and usual requests. The nature of the task and the participant's personality were both considered to be deciding factors. Rossi et al. in [18,19] delved further into which aspects of personality may be necessary: agreeableness, conscientiousness, stable

C. Goutte and X. Zhu (Eds.): Canadian AI 2020, LNAI 12109, pp. 138–150, 2020.
https://doi.org/10.1007/978-3-030-47358-7_13

emotionality all suggested higher disposition for assuming benevolence and thus, a tendency to adopt a trusting stance. Extroverts were also more likely to trust a robot. How trusted agents differ from trusted humans is another theme. De Melo et al. [15,16] indicate that humans experience less guilt when dealing with agent partners than humans, and also reveal that humans are concerned when agents adopt different values during bargaining. That humans are less tolerant of mistakes made by AI agents than by humans has also been observed [25]. These papers draw out some of the differing emotional reactions to AI from users. A final paper that helps to emphasize the importance of considering personalized solutions, when developing AI approaches that consider human acceptance, is that of Anjomshoae et al. [1]. This work draws out the critical value of providing explanations to non-experts but also reveals that few proposals have addressed the significant concern of personalization and context-awareness. Indeed, without sufficient explanation, users may ascribe unfounded mental states to robots that are operating on their behalf [10], so that it is essential to address explainability well. Increased confidence can also arise with sufficient explanation [17], which allows the receiver to perceive the sender's state of mind.

The fact that risk profiles may hold the key to determining how to design our systems is a theme that emerges from the work of Mayer et al. [14] exploring organizational trust. The authors suggest that perceived trustworthiness is comprised of the components of benevolence (belief in wanting good), integrity (adhering to desired principles) and ability (having necessary skills). They raise the point that risk-taking actions may influence future perceptions of trustworthiness.

2 Desiderata for Trusted AI

As outlined in [4], there are a variety of reasons why individuals and organizations may be concerned about employing AI solutions for decision making. These include issues of fairness, explainability, ethics and safety. One question to consider is whether multiple concerns can be addressed simultaneously, with any of the efforts today to adjust our AI solutions in order to engender trust. For instance, are current approaches to address issues of fairness for AI systems in use also helping to increase transparency? We return to discuss the challenge of taking multiple concerns into consideration at once, towards the end of the paper.

The proposal of [4] is to make better use of existing frameworks for modelling agent trustworthiness, precisely in order to compare competing proposals for enabling trusted AI solutions for users. Trust modelling algorithms can identify less reliable sources either by learning from direct experience, together with a prediction of future behaviour or by interpreting reports received from other agents in the environment, judged according to their inherent reliability. The authors sketch how the use of trust modelling, injected into an environment for judging competing explanations from differing sources, can assist in determining which overall consensus is most dependable. This may be effective for settings

such as gauging the value of supervised learning. A specific proposal outlining how to integrate trust modelling into a system aimed at promoting trusted AI is not offered in this short paper. The paper at least argues that not all efforts to address anyone of the desiderata (fairness, accountability, transparency) are equally valuable, and some comparisons are thus necessary.

Our view, however, is that for any attempt to improve outcomes regarding even one of these concerns, there will be a tradeoff. Furthermore, ultimately, user preferences should determine how those tradeoffs are modelled, and thus which competing goal should take precedence. One researcher with the desire to identify user needs for the sake of improved acceptance from users is Kambhampati [12]. References such as [22] suggest that planning and explainability should go hand in hand. Which kinds of explanations provide the best outcomes depend at times on the perceived model differences of the users [26]. There is also an effort to describe problems in a space of plan interpretability [2] where the computational capability of the observer is an issue. Agents can opt to make intentions clear to users or to obfuscate. Explicable planning can be viewed as an effort to minimize the distance from the plan expected by the human as well [2] so that mappings of plan distances need to be estimated and reasoned with. True human-aware planning [22] makes adjustments during the plan generation process itself, effectively beginning to reason about trading off the cost and computation time of plans in a way that serves the human observer best. In essence, the main tradeoff is between sacrificing optimality against the cost of an explanation. What would be especially valuable to consider are different ways in which these sacrifices can be decided based on the particular user at hand and their specific preferences. One relevant dimension that we choose to explore here is that of the user's risk attitude.

3 Reasoning with Differing User Preferences for Trusted AI

In order to support our position that it is valuable for efforts towards trusted AI to consider specific risk preferences of users, we offer three primary arguments. The first two arguments are framed within the context of research which suggests that agents should adopt less optimal but safer plans in scenarios where users opt to observe the plans or the execution of plans of intelligent agents [12]. The first argument opens up the process of requesting observability, clarifying that despite a user's initial risk profile, they may progressively update their preferences about the Agent's plan and its need for explainability. The second argument distinguishes the concept of user trust in an agent's plan and the notion of user's inherent risk profile. We outline how cautious users may still prefer less safe plans under certain circumstances. After discussing how risk profiles can influence trusted AI solutions for explainability, we then briefly explain how fairness decisions may also be influenced by risk profiles of users, as our third argument. We conclude with suggested steps forward to continue to map out user-specific approaches for engendering user trust.

3.1 Game-Theoretic Reasoning

We are considering one especially relevant starting point for framing an avenue to integrate reasoning about user preferences as the detailed proposal of [21], which reflects on the kind of costs that users may need to bear within the setting of a game-theoretic model of trust. Interestingly, the risks associated with robots or humans, making certain decisions with respect to their partnership for executing the real-world plan are discussed within this paper as well. The framework at least suggests that each user may have differing preferences, so that decisions about actions such as observing the Robot or not will incur costs that users may willingly opt to endure, again with certain consequences of doing so with respect to optimality. Borrowing a similar game-theoretic framework, we begin to study how user risk profiles may end up suggesting differing outcomes for the trusted AI effort. Our game-theoretic framework for studying AI behaviour under assumptions of risk is as follows.

We first note that in the model of [21], the Robot has a model of the Human's assessment of the Robot. The Human has an option to monitor the Robot's behaviour and stop execution if needed. Since meta-models of human behaviour may be challenging to learn and interpret, we recast the goals and constraints of each actor based on the risk profiles of the humans and also use the word Agent instead of Robot to represent the AI entity. Finally, we re-imagine the cost of inconvenience for the Human as the risk that the Human takes for allowing the Agent to create and execute the plan. We believe that the success of a plan will depend on the explainability, i.e. how well the Human understands the solution given by the Agent. The Agent should have a cost of explainability. Our framework, outlined below, assumes that the Human arrives with a known general risk profile and that the Agent primarily aims to avoid the cost of not achieving the goal at hand. This formulation thus moves beyond the more vague concept of Human mental model, which [21] assumes to be the basis for the Agent's reasoning. For now, we imagine that explainability incurs an additional computational cost, beyond that of making or executing a plan. Our framework proceeds as follows.

1. Agent is the artificial intelligence agent who has the following properties:
   (a) Agent is uncertain of Human's risk assessment of the Agent but knows that their risk profile of the Human is the space of all possible risk profiles. Thus, it has a perceived risk profile that we denote as $\mathcal{R}^H$.
   (b) Agent has plan π_p. Agent incurs some cost for not achieving the goal.
2. Human is the human actor who requests some information from the Agent.
   (a) Human has a risk profile R^H.
   (b) If Human believes a plan is risky, it can observe and stop at any time. It incurs the cost of observing the partially executed plan and a cost of the inconvenience of not achieving goal.
   (c) Human has non zero cost of observation.
   (d) Human is rational, i.e. they only stop execution if the plan is too risky.

3. Plan
   (a) A plan π_p is a set of sequential decisions made by the agent. π_s is the safest plan that satisfies all of the models of R^H and π_r satisfies none.

The Human has following strategies: Observe, stop execution at that time ($S_{O,\sim E}$); observe at some time but not stop the execution ($S_{O,E}$); not observe and not stop ($S_{\sim O,E}$); and not observe and stop at any time($S_{\sim O,\sim E}$).

The payoffs for the Human and Agent for a plan are given in Table 1. Note there is a negative sign as the actor incurs a penalty equal to the cost. These calculations are based on the following utilities.

1. **Agent**
   (a) Cost of making plan is $C^A_P(\pi_p)$.
   (b) Cost of explaining is $C^A_E(\pi_p)$.
   (c) We can say that $C^A_E(\pi_p) > C^A_P(\pi_p)$ i.e. cost of explaining a plan is greater than cost of making the plan.
   (d) Cost of explaining until a partial plan ($\hat{\pi}_p$) will be less than cost of explaining entire plan (π_p) i.e., $C^A_E(\pi_p) > C^A_E(\hat{\pi}_p)$
   (e) Cost of not achieving goal (G) is $C^A_{\hat{G}}$. We can assume that the safest plan doesn't have a cost of failure i.e., $C^A_{\hat{G}} = 0$ when $\pi_p = \pi_s$.
2. **Human**
   (a) Cost of observing the plan until some plan ($\hat{\pi}_p$) has been executed is $C^H_E(\hat{\pi}_p)$.
   (b) Cost of observing at the end is $C^H_E(\pi_p)$.
   (c) Cost of not achieving goal is $C^H_{\hat{G}}$.
   $C^H_{\hat{G}} = 0$ when $\pi_p = \pi_s$.
   (d) Risk of executing a plan (π_p) is $R^H(\pi_p)$.
   We can assume that there is no risk in the safest plan i.e., $R^H(\pi_s) = 0$.
   (e) Risk of executing a full plan (π_p) is more than that of a partial plan ($\hat{\pi}_p$) i.e. $R^H(\pi_p) > R^H(\hat{\pi}_p)$.

Table 1. Normal form game matrix for our formulation. The top line is the human's payoff and the bottom line is the agent's payoff.

	$S_{O,\sim E}$	$S_{\sim O,\sim E}$	$S_{\sim O,E}$	$S_{O,E}$
π_p	$-C^H_E(\hat{\pi}_p) - C^H_{\hat{G}}$ $-C^A_P(\pi_p) - C^A_{\hat{G}} - C^A_E(\hat{\pi}_p)$	$-C^H_{\hat{G}}$ $-C^A_P(\pi_p) - C^A_{\hat{G}}$	$-R^H(\pi_p)$ $-C^A_P(\pi_p) - C^A_E(\pi_p)$	$-C^H_E(\hat{\pi}_p) - R^H_E(\pi_p - \hat{\pi}_p)$ $-C^A_P(\pi_p) - C^A_E(\pi_p)$
π_s	$-C^H_E(\pi_s)$ $-C^A_P(\pi_s) - C^A_E(\pi_s)$	NA	0 $-C^A_P(\pi_s) - C^A_E(\pi_s)$	$-C^H_E(\pi_s)$ $-C^A_P(\pi_s) - C^A_E(\pi_s)$

Note that, if the Human completely trusts the system ($\pi_p = \pi_s$), then the Nash Equilibrium is when the Human selects $S_{\sim O,E}$, i.e. the plan is executed without observing. If the Human completely distrusts the system ($\pi_p = \pi_r$), then the Nash Equilibrium is when the Human selects $S_{\sim O,\sim E}$, i.e. they do not execute the plan without observing. In the discussion that follows, we explore how

a Human's risk profile can influence the Agent's reasoning (moves in the game), equating the risk profile with certain assumptions about whether to require explainability.

1. **Risk Averse:** In this case, we can assume that $R^H(\pi_p) > C^H_E(\pi_p) + C^H_{\hat{G}}$, i.e. it is riskier to let any plan run its course without observing. The human will prefer either $S_{O,\sim E}$ or $S_{O,E}$. In order for the goal to be achieved, the Agent's costs $C^A_{\hat{G}} > C^A_E(\pi_p) - C^A_P(\hat{\pi}_p)$, i.e. the cost of achieving the goal must at least be greater than the cost of explaining the rest of the task.
2. **Risk Taking:** It is most important to get to the goal $R^H(\pi_p) < C^H_{\hat{G}}$. In this case, the Agent does not need to change its strategy. However, the Agent can make a more explainable plan to start with such that the Human does not need to observe at all.

We begin to sketch one way in which the Agent's decision procedure could be represented. The Agent needs to reason about whether to execute its plan at hand or to adjust its plans based on the user's perceived risk profile. The Agent will focus on explainability at the expense of accuracy (i.e. optimality), with a risk-averse human. It may forgo attempts at explainability to promote accuracy if the user is more risk-seeking. One interesting question is whether the Agent believes it can achieve the kind of explainability desired by the Human. The reasoning could proceed in the following fashion. If the Agent believes that the plan is more difficult to explain than it is worth trying to do so, it could integrate a kind of mixed-initiative dialogue with the user, asking for more direction. The Agent may view this as a scenario where the user's risk threshold may be exceeded, if that Agent proceeds with the plan at hand. The Human could be presented with options to either dismiss the Agent's execution of actions on its behalf (take manual control) or could instruct the Agent to begin reformulating a new plan from the current state (making more of an effort with the explainability). This procedure is intended to give more agency to the Human, and to prevent unintended outcomes that the Human's risk profile suggests should be avoided.

One interesting option is to have the Agent reflect on whether it should continue to execute its plan, each time a new step is taken. Such a decision procedure could be run as follows. At each decision element $p_i \in \pi_p$, the Agent reflects on whether the cost of achieving its goal is more than the cost of explanation. If so, it prompts the user. The user's risk profile can then be progressively learned during the dialogue as well. If the user opts to observe an execution, this can cause the Agent to increase its belief that the user is risk-averse. If the user proceeds for some time without requesting any observations, the Agent may opt to reduce its view of the user as risk-averse.

We view the primary decision-making of the Agent in a cycle of execution as one of assessing $\mathcal{R}^H(\pi)$ as $\mathcal{R}^H(\pi) > C^H_E(\pi) + G^H_{\hat{G}}$ and increasing the cost by $cost = cost + C^A_E(\pi)$. When steps of a plan have begun executing without a request for observability from the user, the Agent can reduce $\mathcal{R}^H$ (its view of the Human's risk profile); likewise, as Humans engage in observing, the Agent may increase its belief that the Human is risk-averse.

3.2 Distinguishing Trust and Risk Averseness

In Sengupta et al.'s [21] model, the human H has three strategies, to only observe the planning process (OP), to only observe the execution process (OE), or to not observe the robot at all (NO-OB). They argue that it is the Human's lack of trust in the Agent, which causes the Agent to opt for a safer plan (abandoning one which may be more accurate but also riskier). We note that while risk averseness of a user can translate into a distinct consideration for an Agent to reason about explaining its actions, it is also valuable to consider distinguishing between the Human's trust of the Agent and their inherent risk profile. In Sengupta et al.'s, original model, the trust boundary of the Human is derived to ensure that the Agent will never execute the risky plan. However, in some situations, the risky plan might be desired by the Human, as it incurs a lower cost. Thus, the Human chooses a mixed strategy $q = [q_P, q_E, (1 - q_P - q_E)]^T$ over the actions OP, OE and NO-OB respectively. For the risky plan to be worth executing, the expected utility given trust boundary q has to be higher for π_p than for π_s by α. $\alpha \in (0, 1]$ and we must consider the cost of execution of the plan $C_E^A(\pi_p)$

$$\begin{aligned} \mathbb{E}_q[U(\pi_s)] &< \mathbb{E}_q[U(\pi_p)] \times \alpha \qquad \text{i.e.} \\ (-C_P^A(\pi_s) - C_E^A(\pi_s)) &< ((-C_P^A(\pi_p) - C_E^A(\pi_p) - C_{\hat{G}}^A) \times q_P \\ &\quad +(-C_P^A(\pi_p) - C_E^A(\hat{\pi}_p) - C_{\hat{G}}^A) \times q_E \\ &\quad +(-C_P^A(\pi_p) - C_E^A(\pi_p)) \times (1 - q_P - q_E)) \times \alpha \end{aligned} \tag{1}$$

If the Agent knows the user's risk profile and trust boundary, using the above equation would enable reasoning about whether the risky plan is worth executing instead of the safe plan. It is important to note, however, that while the user's risk profile (obtained through some initial questionnaire, for instance) is relatively stable, the user's trust boundary (determined by noticing how often they observe the planning process or execution) is constantly changing. If a user observes the Agent planning or executing a risky plan, for instance, then their trust in the Agent may be lowered. If a user observes the Agent executing a safe plan, their trust in the Agent may increase. As such, a progressive update of the model of the Human is necessary for the Agent's decision procedure. But merely relying on a modelling of the Human's trust may cause the Agent to be overly cautious in its planning.

4 Fairness and Explainability

In the previous section, we studied how designing for trusted AI may require reasoning about differing preferences from users with respect to accuracy and explainability. In this section, we delve further into the consideration of fairness, as another pillar of trusted AI for which distinct user risk profiles may suggest alternative designs for reasoning about costs and tradeoffs. We begin by exploring further the kinds of concerns that exist today regarding the fairness of AI systems used for decision making. We reflect on whether solutions for fairness

can also satisfy a desire for explainability. We then discuss how to reason about individual preferences when trying to balance these two considerations. One view is that risk-averse individuals may be willing to accept systems that have been demonstrated to be fair, even if the methods for achieving this fairness cannot be fully explained.

As motivation for this position about tradeoffs for trusted AI, consider the case where an organization is running AI algorithms in order to make decisions on whether to hire a new employee. One might imagine setting the risk profile of the organization initially to be extremely risk-averse with respect to investment in fairness. This means that a solution that has inadequately considered fairness is problematic, as it could result in the organization being charged with discrimination. Consider as well the context where the hiring decisions are derived from modelling various features of successful employees in the past so that the solution is data-driven machine learning. In this case, the data used for training, as well as the reasoning about which features constitute an ideal employee, must both be under the microscope with respect to fairness. Now suppose that cases of clear discrimination do not arise, but that explanations for failing to hire a particular individual are desired, and these are difficult to articulate clearly, as they are tied to some sophisticated deep learning methodology. One might imagine being disappointed in the failure of explainability, but willing to put up with this consequence, if the required fairness has at least been attained. We begin with the first observation that the concern at hand may be with respect to the fairness of the data or with respect to the decision-making algorithm, which is making use of the data that is provided.

There are three major approaches towards achieving fairness in decision-making algorithms. We can modify the input data distribution to reduce bias at the source and thus train our model on the cleaner data. This method is called pre-processing. An alternate approach is to instead regulate the loss function of the classifier by adding fairness measures as regularization terms. This helps to control the tradeoff between fairness and the overall accuracy of the system. [9] showed that machine learning models amplified representation disparity over time and proposed to alter the loss function to minimize the worst-case risk for the minority groups. Notice that in the above approaches, we need to have access to the underlying data, which is not possible in many cases. Thus, in an alternative approach, one may first use the original data to train a classifier and then generate another classifier. This new classifier is independent of the original data and is created using just the original classifier and the protected attribute. However, we then need to ensure that the new classifier is fair by some definition of fairness. This approach is referred to as post-processing. One way to formalize this idea was proposed by [8], where they learn a probability distribution which controls whether to change the value of the predicted output from the original classifier or to keep it the same. The probability distribution is learnt by solving an optimization problem, which ensures that the desired fairness constraints are met while keeping the accuracy close to that of the original classifier.

Even if we have a user who is risk-averse with respect to explainability, there may be different options for addressing their needs. [6] uses crowd-sourcing in the form of Amazon Mechanical Turks to provide insights on how different styles of explanation impact people's fairness judgment of ML systems. They show people certain explanations of a model and ask them to differentiate between global explanations (describing the whole model) and local explanations (justifying a particular decision). They argue that it depends on the kinds of fairness issues and user profiles and that there is no one-size-fits-all solution for an effective explanation. Finally, they show that individuals' prior positions on algorithmic fairness also influence how they react to explanations. They argue for providing personalized forms of explanations to users. The authors do not, however, provide specific insights into how this personalization can be achieved.

Tradeoffs and Alternative Definitions for Fairness

In order to map out a decision procedure to respect user preferences with respect to fairness, the same concerns of Kambhampati and his coauthors [12,21,22], namely tradeoffs between accuracy and explainability, seem to arise. The point is that efforts to examine the success of the classification algorithm, with respect to the data on which it has been trained, may fail to consider bias, so that obtaining what appears to be well-respected performance accuracy, may still reflect critical failure with respect to fairness. A small example here helps to draw out why this might be the case.

Imagine we have a task to determine if a student will be successful in graduate school using some test scores. For simplicity, let us assume that students belong to two demographic groups. Students of one demographic group might be in the majority. If for certain reasons, students belonging to this majority demographics have higher scores (they are rich and give exams multiple times) than the students in the minority group. Therefore a classifier trained to get the best accuracy might more often reject the students belonging to the minority group. Thus, a more accurate solution does not mean a fair solution. Optimizing for average errors fits the majority error. We can obtain Pareto improvements by using group memberships. Current models tend to be inherently designed to be more accurate and not fair. They pick up the bias present in the data and, in some cases, amplify it [27]. [5] showed that if we have to satisfy even a single fairness criterion, we will sacrifice on the utility (accuracy).

We note as well that in some cases, users may not be risk-averse regarding fairness (e.g. they believe that this is not an issue for their particular application), while they may instead be much less forgiving of a failure to explain (e.g. unwilling to run software in their firm which cannot be justified to shareholders). For these individuals, the cost profile has changed and, all of this still needs to be considered against the primary aim of producing a system that provides important overall accuracy.

It is also important to note that there are alternate definitions of fairness, so that any effort to achieve this aim for trusted AI, may need to be attuned to a more precise measurement. Three widely used ways of modelling fairness are disparate impact, individual fairness or equalized odds. It may, therefore, be

important to determine which of these considerations is paramount for the user, for their given context, before proceeding to reason about how best to design and where the tradeoffs between costs are best determined. Our discussion below also points out some challenges in achieving each of these differing perspectives, to date. Disparate Impact is measured by $Pr(Y = 1|A = 0)/Pr(Y = 1/A = 1) <= t$, where t is a threshold value, Y represents an outcome (1 is a positive event) and A represents the protected attribute (0 is the minority/protected class). Thus, it means that the probability of the classification event (getting admitted) should be independent of the protected attribute (race). For our scenario of admission to graduate school, getting admitted should, for instance, be independent of race. This measure may not ensure fairness, however, as the notion permits that we accept qualified applicants in the demographic $A = 1$ but unqualified individuals in $A = 0$, as long as the percentages of acceptances match. In addition, demographic parity may cripple utility in cases where the target variable Y is correlated with A. Individual Fairness [7] defines the notion that similar individuals should be treated similarly. How best to frame this metric is currently unresolved. Equalized Odds measure was defined by [8]. The constraint requires that the classifier(represented by Y) has equal true (y=1) and false (y=0) positive rates across the two demographics A=0 and A=1. $Pr(Y = 1|A = 0, C = y) = Pr(Y = 1|A = 1, C = y), y \in 0, 1$ C represents the true label present in the dataset. However, equalized odds enforces that the accuracy is equally high in all demographics, punishing models that perform well only on the majority.

5 Acquiring and Updating User Risk Profiles

A question to resolve is how best to represent a user's risk profile. A broad classification of the user as risk-averse or risk-seeking would enable an initial step forward with the reasoning proposed in this paper, namely to vary the outcomes of algorithms aiming to achieve trusted AI. We argue that it is desirable to make these adjustments. Any effort to model a user requires decisions along several fronts per the longstanding field of user modeling [13]: representing the user (what to model and how to represent), deciding when to update the user model, deciding how to reason with the user model to adjust an intelligent agent's decision making and how to acquire the user model. It is only the first of these elements that we have focused on so far.

Some research has explored the methods best able to elicit user risk profiles [3]. These authors mention a suggestion that individuals attempt to maximize some specific utility functions, as well. They ultimately caution against using conclusions within one limited context in order to predict what users will prefer in other scenarios. The fact that users may deviate from expected utility theory is also mentioned in the work of [11]. These authors then suggest that solutions for eliciting preferences should be able to function well, for cases where users are instead of making decisions based on cumulative prospect theory [24]. One element sketched in this paper is an expected minimax regret (EMR) heuristic,

leading to the selection of queries for the user based on maintaining the lowest expected pairwise maximum regret, between pairs of possible decisions. It is therefore quite important to acknowledge the challenges of acquiring an effective representation of a user's risk profile, in order to then reason with that information when making higher-level decisions about how best to respect the various elements of trusted AI. One critical point is the fact that user risk profiles are not static, but may dynamically change as the user can experience various outcomes. Moreover, user risk profiles may also be quite varied, depending on the specific context. This point has been acknowledged well already in the multiagent trust modelling community, where trying to engender trust in a user may need to differ considerably, depending on which features are most important to that user (e.g. quality or cost, for e-commerce transactions) [23]. While the work above reinforces the point that it is important to develop complex strategies for properly determining the risk profiles of users, it is still the case that if some risk attitudes were known, AI system choices could be adjusted to increase acceptability from users.

6 Conclusion and Future Work

This paper proposes new directions for addressing trustworthiness of AI, presenting a particular viewpoint for designing future AI systems. We reveal that there are tradeoffs when aiming for trusted AI, that user risk profiles matter and can be integrated into decisions about how to design our systems, and that we should be considering solutions where different costs are more central, depending on the user. We have also reflected on basic considerations of accuracy, explainability and fairness, revealing that distinct needs and definitions may be at play; we have also sketched the process for reasoning about costs and risk profiles. While we comment on how broad, straightforward classification of user risk profiles enable an initial solution, we also acknowledge that this consideration is considerably more complex, requiring a collection of more complex reasoning processes.

The most obvious first step for future work is to expand upon representing user risk aversion, reflecting further on how these profiles are best known or acquired, and allowing finer-grained distinctions (for example, enabling users to be risk-averse in certain specified contexts, and more forgiving for other tasks being executed by the AI system). We have sketched only one small proposal for designing AI systems based on risk profiles and costs; ours is embedded in the game-theoretic analysis. Future work should proceed to calibrate gains for trusted AI using our approach. We should also consider many other desiderata of AI systems and continue to determine how decisions made by AI systems can be modulated by these considerations in a way that is faithful to individual user needs. We also acknowledge that reasoning more broadly about user preferences rather than risk profiles per se might open up a deeper set of approaches. But our position is that focusing on risk aversion as the critical element which may require sacrificing accuracy for explainability (or other trusted AI concerns) is

a very powerful and effective stand-in,[1] one that shows promise as we continue our dialogue with those invested in securing better acceptance of AI from non-experts.

References

1. Anjomshoae, S., Främling, K., Najjar, A.: Explanations of black-box model predictions by contextual importance and utility. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) EXTRAAMAS 2019. LNCS (LNAI), vol. 11763, pp. 95–109. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30391-4_6
2. Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D.E., Kambhampati, S.: Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In: Proceedings of the International Conference on Automated Planning and Scheduling, vol. 29, pp. 86–96 (2019)
3. Charness, G., Gneezy, U., Imas, A.: Experimental methods: eliciting risk preferences. J. Econ. Behav. Organ. **87**, 43–51 (2013)
4. Cohen, R., Schaekermann, M., Liu, S., Cormier, M.: Trusted AI and the contribution of trust modeling in multiagent systems. In: Proceedings of AAMAS, pp. 1644–1648 (2019)
5. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806. ACM (2017)
6. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275–285. ACM (2019)
7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226. ACM (2012)
8. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, pp. 3315–3323 (2016)
9. Hashimoto, T.B., Srivastava, M., Namkoong, H., Liang, P.: Fairness without demographics in repeated loss minimization. arXiv preprint arXiv:1806.08010 (2018)
10. Hellström, T., Bensch, S.: Understandable robots-what, why, and how. Paladyn J. Behav. Robot. **9**(1), 110–123 (2018)
11. Hines, G., Larson, K.: Preference elicitation for risky prospects. In: Proceedings of AAMAS, pp. 889–896 (2010)
12. Kambhampati, S.: Synthesizing explainable behavior for human-ai collaboration. In: Proceedings of AAMAS. Richland, SC, pp. 1–2 (2019)
13. Kass, R., Finin, T.: Modeling the user in natural language systems. Comput. Linguist. **14**(3), 5–22 (1988)
14. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**(3), 709–734 (1995)

[1] In contrast with the more general term of human mental models proposed in [12].

15. Melo, C.D., Marsella, S., Gratch, J.: People do not feel guilty about exploiting machines. ACM Trans. Comput. Hum. Interac. (TOCHI) **23**(2), 8 (2016)
16. de Melo, C.M., Marsella, S., Gratch, J.: Do as I say, not as I do: challenges in delegating decisions to automated agents. In: Proceedings of AAMAS, pp. 949–956 (2016)
17. Nomura, T., Kawakami, K.: Relationships between robot's self-disclosures and human's anxiety toward robots. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-vol. 03, pp. 66–69. IEEE Computer Society (2011)
18. Rossi, A., Dautenhahn, K., Koay, K.L., Walters, M.L.: The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. Paladyn J. Behav. Robot. **9**(1), 137–154 (2018)
19. Rossi, A., Holthaus, P., Dautenhahn, K., Koay, K.L., Walters, M.L.: Getting to know pepper: effects of people's awareness of a robot's capabilities on their trust in the robot. In: Proceedings of the 6th International Conference on Human-Agent Interaction, pp. 246–252. ACM (2018)
20. Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Would you trust a (faulty) robot?: effects of error, task type and personality on human-robot cooperation and trust. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 141–148. ACM (2015)
21. Sengupta, S., Zahedi, Z., Kambhampati, S.: To monitor or to trust: observing robot's behavior based on a game-theoretic model of trust. In: Proceedings of the Trust Workshop at AAMAS (2019)
22. Sreedharan, S., Kambhampati, S., et al.: Balancing explicability and explanation in human-aware planning. In: 2017 AAAI Fall Symposium Series (2017)
23. Tran, T.T., Cohen, R., Langlois, E., Kates, P.: Establishing trust in multiagent environments: realizing the comprehensive trust management dream. TRUST@ AAMAS **1740**, 35–43 (2014)
24. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of uncertainty. J. Risk Uncertainty **5**(4), 297–323 (1992)
25. Yuksel, B.F., Collisson, P., Czerwinski, M.: Brains or beauty: how to engender trust in user-agent interactions. ACM Trans. Internet Technol. (TOIT) **17**(1), 2 (2017)
26. Zahedi, Z., Olmo, A., Chakraborti, T., Sreedharan, S., Kambhampati, S.: Towards understanding user preferences for explanation types in model reconciliation. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 648–649. IEEE (2019)
27. Zhao, J., et al.: Men also like shopping: reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017)