Artificial Intelligence and Law
https://doi.org/10.1007/s10506-024-09393-y

ORIGINAL RESEARCH

A comparative user study of human predictions in algorithm-supported recidivism risk assessment

Manuel Portela[1] · Carlos Castillo[1,2] · Songül Tolan[3] · Marzieh Karimi-Haghighi[1] · Antonio Andres Pueyo[4]

Accepted: 11 February 2024

Abstract

In this paper, we study the effects of using an algorithm-based risk assessment instrument (RAI) to support the prediction of risk of violent recidivism upon release. The instrument we used is a machine learning version of RiskCanvi used by the Justice Department of *Catalonia, Spain*. It was hypothesized that people can improve their performance on defining the risk of recidivism when assisted with a RAI. Also, that professionals can perform better than non-experts on the domain. Participants had to predict whether a person who has been released from prison will commit a new crime leading to re-incarceration, within the next two years. This user study is done with (1) *general* participants from diverse backgrounds recruited through a crowdsourcing platform, (2) *targeted* participants who are students and practitioners of data science, criminology, or social work and professionals who work with RisCanvi. We also run focus groups with participants of the *targeted* study, including people who use *RisCanvi* in a professional capacity, to interpret the quantitative results. Among other findings, we observe that algorithmic support systematically leads to more accurate predictions from all participants, but that statistically significant gains are only seen in the performance of *targeted* participants with respect to that of crowdsourced participants. Among other comments, professional participants indicate that they would not foresee using a fully-automated system in criminal risk assessment, but do consider it valuable for training, standardization, and to fine-tune or double-check their predictions on particularly difficult cases. We found that the revised prediction by using a RAI increases the performance of all groups, while professionals show a better performance in general. And, a RAI can be considered for extending professional capacities and skills along their careers.

Keywords Recidivism · Automated decision-making · Risk assessment instrument · Human oversight · Machine learning

Carlos Castillo, Songül Tolan, Marzieh Karimi-Haghighi and Antonio Andres Pueyo have contributed equally to this work. Songül Tolan has left the JRC and now works for the German Federal Ministry of Labour and Social Affairs.

Extended author information available on the last page of the article

Published online: 15 March 2024

1 Introduction

Since the 1970s the use of Risk Assessment Instruments (RAI) in high stakes contexts such as medicine or criminal justice, together with their risks and benefits, have been a subject of debate across various disciplines. RAIs may increase the accuracy, robustness, and efficiency in decision making (Kleinberg et al. 2018); however, they can also lead to biased decisions and, consequently, to discriminatory outcomes (Angwin et al. 2016; Skeem et al. 2016). Understanding the performance of a RAI requires looking beyond the statistical properties of a predictive algorithm, and considering the quality and reliability of the decisions made by humans using the RAI (Green 2021). This is because high-stakes decisions are rarely made by algorithms alone, and humans are almost invariably 'in-the-loop', i.e., involved to some extent in the decision making process (Binns and Veale 2021). Indeed, the General Data Protection Regulation (GDPR)[1] in Europe gives data subjects the right 'not to be subject to a decision based solely on automated processing' (Article 22), and the proposed Artificial Intelligence Act[2] in its draft published in April 2021 and approved by the European Parliament on 2024, considers criminal risk assessment a 'high-risk' application subject to stringent human oversight.

Our work involves a sequence of studies outlined in Fig. 1 and described in the next sections. We develop and test different user interfaces of a machine learning version of RisCanvi, the main RAI used by *Catalonia*'s criminal justice system. We ask participants to predict the re-incarceration risk[3] based on the same factors used

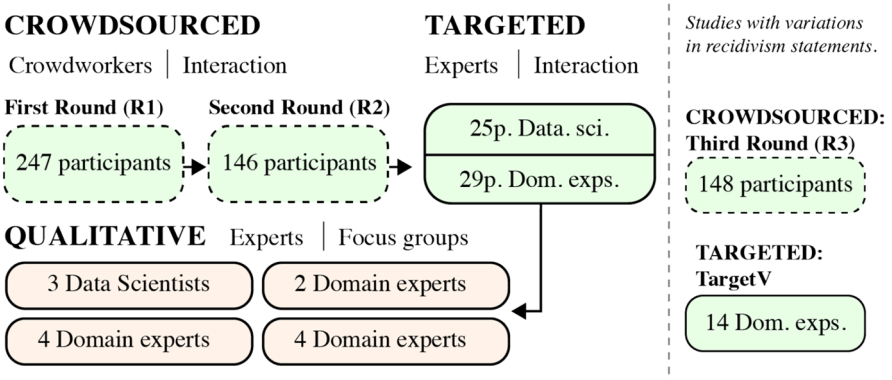


Fig. 1 Sequence of studies and number of participants. The figure shows the three experimental studies (one targeted and two crowdsourced) and the different focus groups made during our study. Additional studies were made after by changing the statements about violent recidivism

[1] Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679. Accessed Jan 2022.

[2] Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Accessed Jan 2022.

[3] Note: Re-arrest and re-incarceration are not necessarily a good proxy for re-offense, our dataset contains information about re-incarceration.

by RisCanvi, such as criminal history, which empirically affect the recidivism risk of individuals. Some participants are additionally shown the risk that the RAI predicts using the same factors. Our primary goal is to assess how the interaction with the studied RAI affects human predictions, their accuracy, and their willingness to rely on a RAI for this task.

As most previous studies on this topic, we partially rely on crowdsourced participants (Dressel and Farid 2018; Grgic-Hlaca et al. 2019; Green and Chen 2020; Fogliato et al. 2021). Controlled in-lab/survey experiments and crowdsourced experiments have the limitation that participants do not face the real world consequences that professional decisions have on the lives of inmates. In addition, untrained crowdworkers may exhibit different decision making behaviour than trained professionals. The former limitation can only be addressed through studies that analyze the real-world adoption of a RAI through observational methods (Berk 2017; Stevenson 2018; Stevenson and Doleac 2021). However, these studies usually face the difficulty of isolating the effect of RAI adoption from other changes that co-occur in the study period. The latter can be addressed in an experimental setting by recruiting professional participants, as we do in this paper. To the best of our knowledge, most studies focus on crowdsourced participants. This might be the first study that, in addition to a crowdsourced study, runs a *targeted* study which results are supported with a validation from a focus group. We recruited students and professionals of data science as well as domain experts (with a background in criminology and social work), including people who work within *Catalonia*'s criminal justice system and use *RisCanvi* in a professional capacity. Finally, we conducted a *qualitative study* with small sub-groups of the *targeted* user study, particularly professionals within the Justice Department of *Catalonia*, as well as data scientists. Our main contributions are:

- We confirm previous results that show how accuracy in decision making slightly improves with algorithmic support, how participants adjust their own predictions in the direction of the algorithmic predictions, and how different scales in risk communication yield different levels of accuracy.
- We describe differences between targeted participants and crowdsourced workers. Despite identical experimental conditions and tasks, we find that the predictions differ between these groups and *targeted* participants outperform crowdsourced participants in terms of accuracy.
- We provide insights into how professionals use RAIs in real-world applications from our focus groups. Our interviewees would not foresee using a fully automated system in criminal risk-assessment, but they see benefits in using algorithmic support for training and standardization, and for fine-tuning and double-checking particularly difficult cases.

The remainder of this paper is structured as follows. First, we give an overview of related work (Sect. 2). Next, we describe our approach, including the variables of our study (Sect. 3), as well as the materials and procedures we employ (Sect. 4). We present the experiment setup for crowdsourced and targeted participants (Sect. 5), and the obtained results from both groups (Sect. 6). Then we present the results

from the focus groups (Sect. 7). Finally, we discuss our findings (Sect. 8), as well as the limitations of this study and possible directions for future work (Sect. 9).

2 Related work

2.1 Risk assessment instruments (RAI) for criminal recidivism

Law enforcement systems increasingly use statistical algorithms, e.g., methods that predict the risk of arrestees to re-offend, to support their decision making (Goel et al. 2019; Chiusi et al. 2020). RAIs for criminal recidivism risk are in use in various countries including Austria (Rettenberger et al. 2010), Canada (Kröner et al. 2007), Germany (Dahle et al. 2014), Spain (Andrés-Pueyo et al. 2018), the U.K. (Howard and Dixon 2012), and the U.S. (Desmarais and Singh 2013). There are ethical and legal aspects to consider, as algorithms may exhibit biases, which are sometimes inherited from the data on which they are trained (Barocas and Selbst 2016). However, some argue that RAIs bear the potential for considerable welfare gains (Kleinberg et al. 2018). The literature shows that decisions based on RAIs' scores are never made by an algorithm alone. Decisions in criminal justice are made by professionals (e.g., judges or case workers) (Bao et al. 2021), sometimes using RAIs (Stevenson and Doleac 2021). Consequently, algorithms aimed at supporting decision processes, especially in high-risk contexts such as criminal justice, cannot be developed without taking into account the influences that institutional, behavioural, and social aspects have on the decisions (Selbst et al. 2019). Furthermore, human factors such as biases, preferences and deviating objectives can also influence the effectiveness of algorithm-supported decision making (Jahanbakhsh et al. 2020; Mallari et al. 2020). Experienced decision makers may be more inclined to deviate from an algorithmic recommendation, relying more on their own cognitive processes (Green and Chen 2020). Moreover, trained professionals, such as probation officers, may prefer to rely on their own decision and not just on a single numerical RAI prediction. Any additional information that they consider may be used as a reason to deviate from what a RAI might recommend for a case (McCallum et al. 2017). There are other reasons why humans disagree with an algorithmic recommendation. For instance, the human's objectives might be misaligned with the objective for which the algorithm is optimized (Green 2020), or the context may create incentives for the human decision maker not to follow the algorithm's recommendation (Stevenson and Doleac 2021). Sometimes humans are unable to evaluate the performance of themselves or the risk assessment, and engage in 'disparate interactions' reproducing biased predictions by the algorithm (Green and Chen 2019). Another reason could be algorithm aversion, e.g., human decision makers may discontinue the use of an algorithm after observing a mistake, even if the algorithm is on average more accurate than them (Dietvorst et al. 2015; Burton et al. 2020). In contrast, controlled user studies in criminal risk assessment indicate that crowdsourced participants tend to exhibit *automation bias*, i.e., a tendency to over-rely on the algorithm's prediction (Dressel and Farid 2018; Bansak 2019).

Effective human-algorithm interaction depends on users' training with the tool, on the experience of the human decision maker with the algorithm, and on the specific professional domain in which the decision is made. Therefore, some researchers have studied the impact of the adoption of RAIs in criminal justice decision-making in real-world applications (Berk 2017; Stevenson 2018; Stevenson and Doleac 2021). These observational studies yield valuable insights, but the conditions of adoption as well as the design of the RAI cannot be controlled, making it difficult to isolate the effect of the RAI on the studied outcome.

2.2 Controlled user studies and interaction design of RAIs

Algorithm-supported human decision making has also been studied in controlled experiments (Dressel and Farid 2018; Green and Chen 2019, 2019; Grgić-Hlača et al. 2019; Lin et al. 2020; Fogliato et al. 2021). Among these, an influential study by Dressel and Farid in 2018 (Dressel and Farid 2018), showed how crowdsourced users recruited from Amazon Mechanical Turk (AMT) were able to outperform the predictions of COMPAS, a RAI that has been subject to significant scrutiny since the seminal work of Angwin et al. (2016). Follow-up studies criticized Dressel and Farid's study, noting that participants were shown the ground truth of each case (i.e., whether or not the person actually recidivated) immediately after every prediction they make, which does not correspond to how these instruments are used in practice. Without this feedback, human predictions that were not supported by algorithms performed worse than the algorithm under analysis (Lin et al. 2020).

The way risk assessments are communicated and integrated in the decision process plays a crucial role in the quality of the predictions. For instance, criminal forensics clinicians have a preference for (non-numerical) categorical statements (such as 'low risk' and 'high risk') over numerical risk levels. However, an experimental survey showed that a RAI providing numerical information elicits better predictive accuracy than if categorical risk levels are used (Zoe Hilton et al. 2008). One issue with categorical expressions is that professionals tend to disagree about the limits of the categories and how these categories represent different numerical risk estimations (Hilton et al. 2015). However, numerical expressions introduce other challenges. For instance, participants in a study perceived violence risk as higher when the risk was presented in a frequency format instead of a percentage format (Hilton et al. 2015). Another question is whether numerical risks should be presented on an absolute or a relative scale. A study with clinicians showed that participants hardly distinguish between absolute probability of violence and comparative risk (Zoe Hilton et al. 2008). Furthermore, besides showing only risk levels, risk assessments could include additional information about the nature of the crime, the factors of the RAI and other factors that may have preventive effects on future re-offense (Heilbrun et al. 1999). Complementary and graphical information can improve the understanding of risk evaluations (Hilton 2017). However, it can also increase the overestimation of risk factors while ignoring other contextual information (Batastini 2019). Nevertheless, the use of different visualization methods is mainly unexplored.

Given the experience from previous work, we build our user interface to test and measure the performance of participants using different categorical risk levels and numerical expressions for risk, specifically absolute and relative risk scales. We conduct a recidivism prediction experiment with crowdsourced participants, but also complement it with *targeted participants*. One of the main novelties of our study resides in assessing how *targeted* participants, including domain experts and data scientists, perform differently than crowdsourced participants. Another key difference of this study with respect to previous work (Dressel and Farid 2018; Green and Chen 2019, 2019; Grgić-Hlača et al. 2019; Lin et al. 2020; Fogliato et al. 2021), is to include experts and conduct focus groups to validate our findings, and to understand their rationale throughout their decision-making process. Additionally, focus groups and interviews with professionals provided valuable insights into how RAIs are perceived and used in practice.

3 Approach and research questions

This paper takes an experimental approach. Participants in our experiments are asked to determine the probability that an inmate will be re-arrested, based on a list of criminologically relevant characteristics of the case. We focus on three main outcome variables (Sect. 3.1): the accuracy of predictions, the changes that participants make to their predictions when given the chance to revise them after seeing the RAI's recommendation, and their willingness to rely on RAIs. The main independent variables (Sect. 3.2) are the background of the participants, and the type of risk scale used. Our research questions (Sect. 3.3) are about the interaction of these variables.

3.1 Outcome variables

3.1.1 Predictive accuracy

The performance of predictive tools including RAIs is often evaluated in terms of the extent to which they lead to correct predictions. Due to the inherent class imbalance in this domain, as most people do not recidivate, most studies (e.g., experimental (Dressel and Farid 2018; Harris et al. 2015; Green and Chen 2019)) do not use the metric *accuracy*, which is the probability of issuing a correct prediction. Instead, it is more common to measure the area under the receiver operating characteristic (**AUC-ROC** or simply **AUC**). The AUC can be interpreted as the probability that a randomly drawn recidivist obtains a higher score than a randomly drawn non-recidivist.

3.1.2 Prediction alignment with the RAI

In this work, we observe users' reliance on the algorithmic support system indirectly by looking at changes in their predictions after observing an algorithmic prediction.

We assume that if users change their initial predictions to align them with those of a RAI, they are implicitly signaling more reliance on that RAI than their initial prediction. In general, the extent to which people are willing to trust and rely on a computer system is related to people's engagement and confidence in it van Maanen et al. (2007), Chancey et al. (2017), Lee and See (2004), and in the case of predictive algorithms, to their perceived and actual accuracy (Yin et al. 2019). Different types of information disclosure can elicit different levels of trust and reliance (Du et al. 2019). Performing joint decisions, i.e., being the human in the loop (De-Arteaga et al. 2020), can increase willingness to rely on a system (Zhang et al. 2020).

3.1.3 Preferred level of automation

The experience of interacting with an algorithm-based RAI may also affect the acceptability of similar algorithms in the future. Algorithm-based RAIs may operate in ways that differ by their *level of automation* (Cummings 2004). At the lowest level of automation, the human makes all decisions completely disregarding the RAI; at the highest level of automation, the RAI makes all decisions without human intervention; intermediate levels represent various types of automated interventions. In general, the level of automation chosen by a user should be proportionate to the performance of the automated system. Both *algorithm aversion* (Burton et al. 2020) or under-reliance, as well as *automation bias* (Mosier et al. 1998) or over-reliance, negatively affect the predictive accuracy of users.

3.2 Participant groups and conditions

In this section we describe the main independent variables that we tested in the experiments.

3.2.1 Participant's educational and professional background

Most user studies on recidivism risk prediction rely on crowdsourced participants from online platforms. The background of participants may change the way they interact with a RAI. Data scientists and statisticians have training on statistics, probability, and predictive instruments. Domain experts with a background in psychology, criminology, or who work within the prison system, have a deeper knowledge of factors driving criminal recidivism. Additionally, domain experts who use RAIs receive training on their usage, and they often have a fair amount of training in applied statistics.

Naturally, in real-world applications case worker decisions are far more consequential than the consequences faced by crowdworkers in their lab-like decision scenarios. Similar to previous work (Green and Chen 2019; Cheng et al. 2019; Yu et al. 2020; Dressel and Farid 2018), we add an incentive (in the form of a bonus payment) for correct predictions in the crowdsourced studies. However, this is to encourage appropriate effort, not to simulate a high-stakes scenario.

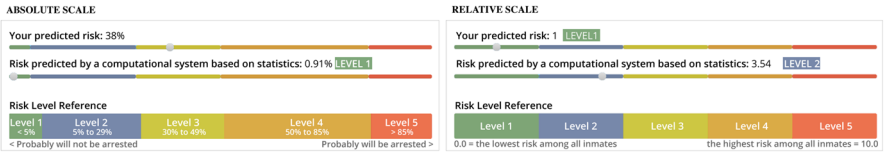


Fig. 2 Risk scales used in our experiments (left: absolute scale, right: relative scale)

We consider three participant groups: (1) crowdsourced workers from unspecified backgrounds, (2) students and practitioners of data science, and (3) students of criminology and people with expertise in the prison system. Recruitment procedures are described in Sect. 4.3.

3.2.2 Risk scales

The literature on risk communication suggests that both numerical and categorical information are useful for different purposes (Zoe Hilton et al. 2008; Jung et al. 2013; McCallum et al. 2017; Storey et al. 2015). Categories alone can be misleading when similar cases are assigned to different categories despite only small differences in their risk (Jung et al. 2013). In our research, we initially used only a categorical scale,[4] but then switched to scales that combine both categorical and numerical values; further, we test two different types of numerical scales. The first scale is based on the probability of recidivism, which we denote 'absolute scale' as it expresses a probability in absolute terms. The second scale we use is based on quantiles of the risk distribution in the data, and we call it the 'relative scale' since it is relative to the risk levels of other cases in the data. We also use five categories, for easier comparison with the absolute scale.

Both scales are depicted in Fig. 2. Other elements in that figure are discussed in the following sections.

3.2.3 Additional variables

Many additional variables could have been included but we were mindful of survey length and wanted to minimize survey drop-out. We included three additional variables: numeracy, decision-making style, and current emotional state. Numeracy is the ability to understand and manage numerical expressions. The decision confidence and the type of information that professionals rely on when using RAIs depends on their numerical proficiency (Scurich 2015). Ideally, professionals working with RAIs should have a fairly high level of numerical literacy, as interpreting RAIs requires the understanding of probabilities, which is not common knowledge. Other factors that have been shown to affect people's decision making behaviour are

[4] We group probabilities using a five-level, empirically-grounded recommendation developed by the US Department of Justice and the US National Reentry Resource Center (Hanson et al. 2017).

their decision-making style and current emotional state (Beale and Peter 2008; Lee and Selart 2012).

3.3 Research questions

Based on the variables we have presented, we pose the following research questions:

- **RQ1: Under which conditions do participants using a RAI to predict recidivism achieve the highest predictive accuracy?**
- **RQ2: To what extent do participants rely on the RAI to predict recidivism?**

4 Materials and methods

In this section, we describe the materials (Sect. 4.1) for our user study which consist of a risk prediction instrument based on *RisCanvi* (Sect. 4.1.1) and a selection of cases used for assessment (Sect. 4.1.2). Next, we present a description of the procedure followed by participants (Sect. 4.2), and the way in which they were recruited (Sect. 4.3).

4.1 Materials

4.1.1 *RisCanvi*

This is one of several risk assessment tools used by the Justice Department of *Catalonia* since 2010 (Andres Pueyo, Arbach-Lucioni and Redondo 2018). This tool is applied multiple times during an inmate's time in prison; in most cases, once every six months.

RisCanvi consists of 43 items that are completed by professionals based on an inmate's record and suitable interviews. Then, a team of professionals (with some overlaps with the various interviewers) makes a decision based on the values of the items and the output of RisCanv's algorithm. RisCanvi's algorithm predicts the risks of four different outcomes: committing further violent offenses (violent recidivism), violence in the prison facilities to other inmates or prison staff, self-injury, and breaking of prison permits.

We focus on *violent recidivism*, which is computed based on 23 of the 43 risk factors as used in *RisCanvi*, including criminal/penitentiary record, biographical factors, family/social factors, clinical factors, and attitude/personality factors.

The original *RisCanvi* uses integer coefficients determined by a group of experts; instead, we use a predictor of violent recidivism created using logistic regression that has a better AUC (0.76) than the original *RisCanvi* (0.72) and is more accurate than models created using other ML methods such as random forests or neural networks (Karimi-Haghighi and Castillo 2021). This is done to reduce any effects

of potential shortcomings of *RisCanvi* originating from using hand-picked integer coefficients, instead using a state-of-the-art predictor based on the same items. In consultation with *RisCanvi* creators, we kept exactly the same features it uses, which are not only the result of a statistical analysis, but also of recommendations from a group of experts formed when *RisCanvi* was being developed. When training the logistic regression, we observed the effect of model multiplicity, where multiple algorithms have similar accuracy (Black et al. 2022).

4.1.2 Cases

In this study, we use a dataset of cases used in previous work (Karimi-Haghighi and Castillo 2021). It consists of the last *RisCanvi* protocol items for the inmates released between 2010 and 2013, and for which recidivism was evaluated by the Department of Justice of *Catalonia*. Upon recommendation of our Ethics Review Board, we do not show participants the data of any individual inmate, but instead created semi-synthetic cases using a cross-over of cases having similar features and similar risk levels (for details, see Appendix 2).

We selected 14 cases which contain a mixture of recidivists and non-recidivists, combining cases in which the majority of humans make correct predictions and cases in which they tend to err, and cases in which the algorithm makes a correct prediction and cases in which it errs. In our first crowdsourcing experiment (referred to as R1 in the following) we observed that these cases were not representative of the performance of the algorithm on the overall dataset. Hence, for the second crowdsourcing experiment (R2 in the following) we exchanged 2 cases to bring the AUC from 0.61 to 0.75 which is closer to the AUC of the algorithm on the original data (0.76). Out of the 14 cases, 3 were used as examples during the 'training' phase of the experiments, while participants were asked to predict recidivism for the remaining 11 cases. All participants evaluate the same 11 cases, but in randomized order.

4.2 Procedure

The study obtained the approval of our university's Ethics Review Board in December, 2020. All user studies were conducted between December, 2020 and July, 2021, and done remotely due to the pandemic caused by the SARS-COVID-19 virus. The survey is designed to be completed in less than 30 min and used an interface hosted in our university's server created using standard web technologies (Python and Flask). The survey is structured as follows:

4.2.1 Landing page and consent form

The recruitment (4.3) leads potential participants from different groups to different landing pages, which record which group the participant belongs to. There, participants learn about the research and we ask for their explicit consent for participating.

4.2.2 Demographics and additional variables

Consenting participants are asked three *optional* demographic questions: age (range), gender, and educational level. Then, three sets of questions are asked to capture the following additional variables (described in Sect. 3.2.3):

- *Numeracy*: We use a test by Lipkus et al. (2001), which has been used in previous work (Hilton 2017). It consists of three questions about probabilities, proportions, and percentages, such as 'If a fair dice is rolled 1,000 times, how many times it will come even (2, 4, or 6)?' (Answer: 500). We measure 'numeracy' as the number of correct answers (0 to 3).

- *Decision making style*: The General Decision Making Style (GDMS) (Scott and Bruce 1995) is a well known survey that identifies five types of individual decision making style: rational, intuitive, dependent, avoidant, and spontaneous.

- *Current emotional state*: We used a Visual Analogue Scale (VAS) to account for 7 attitudes (happiness, sadness, anger, surprise, anxiety, tranquility, and vigor). This survey has been used in previous work (Portela and Granell-canut 2017).

4.2.3 Past experience and attitudes towards RAIs

Participants are asked about their knowledge about and experience with RAIs, as well as what they consider as the three most determining features to predict recidivism, out of the ones used by RisCanvi. The final question of this part is about the level of automation they would prefer for determining the risk of recidivism (see Appendix 1).

4.2.4 Training

The training part consists of the risk assessment of three cases (two non-recidivists and one recidivist). The purpose of this part is to prepare participants for the actual evaluation part and to calibrate their assessment to a ground truth reference. Therefore, unlike the actual risk assessments of the evaluation tasks, participants are shown the ground truth (recidivism or no-recidivism) after each one.

4.2.5 Evaluation tasks

The evaluation tasks are the core part of the study and ask participants to predict the probability of violent recidivism for eleven cases. Participants see a list of 23 items that are used by *RisCanvi* to predict violent recidivism (see Appendix 3 for an illustrated reference), and they are asked to select a number, which can be a recidivism probability or a risk level, depending on the condition (see Fig. 2). Additionally, they are asked to select from the list of items the three items that they considered most important in their evaluation, and to indicate their confidence with their prediction on a 5-points scale.

Table 1 Demographics by study group

		Crowd. (R1) N = 247	Crowd. (R2) N = 146	Crowd. (R3) N = 148	TG Dom. Exp N = 29	TG Data. Sci N = 25	TV Dom. Exp. N = 14
Gender	Male	61.0%	54.3%	50.0%	20.7%	40.0%	41.2%
	Female	48.4%	45.7%	47.3%	79.3%	56.0%	58.8%
	Other	0.6%	0.0%	2.7%	0.0%	4.0%	0.0%
Education	Secondary	19.3%	15.5%	11.5%	3.5%	20.0%	5.8%
	Undergraduate	57.5%	68.2%	62.8%	48.2%	56.0%	47.1%
	Postgraduate	22.3%	15.5%	24.3%	48.3%	24.0%	47.1%
Age	18–25	44.3%	53.5%	49.3%	55.2%	76.0%	11.8%
	26–33	25.2%	24.8%	27.0%	10.3%	20.0%	17.7%
	34–45	19.6%	12.4%	14.2%	13.8%	4.0%	23.5%
	45–75	10.9%	9.2%	9.5%	20.7%	0.0%	47.0%
Numeracy	0 (lowest)	11.2%	18.5%	9.4%	17.2%	4.0%	11.8%
	1	15.0%	20.5%	16.2%	24.1%	0.0%	11.8%
	2	24.0%	22.0%	20.3%	6.9%	28.0%	23.5%
	3 (highest)	49.8%	39.0%	54.1%	51.7%	68.0%	52.9%

Participants in the control group are shown just one screen per case to enter their prediction, while participants in a treatment group are shown a second screen for each case, displaying the algorithm's prediction. This second screen also shows participants their initial prediction for comparison, and allows them to optionally change it. In both screens, participants indicate the confidence in their prediction before continuing.

4.2.6 Closing survey

The experiment ends with a final questionnaire and an evaluation of the entire process. This questionnaire repeats some of the questions made in the beginning, such as the preferred level of automation, the emotional state, and the three features they consider most important in predicting recidivism. Additionally, participants can leave a comment or feedback about the study.

4.3 Participant recruitment

A summary of the participants' demographics is shown in Table 1. The crowdsourced study consisted of three rounds (**R1**, **R2** and **R3**) for which we recruited participants via Prolific.[5] We selected residents of *Catalonia*, between 18 to 75 years

[5] Prolific is a crowdsourcing platform specialized in supporting scientific research. It is available at: https://www.prolific.co/

Table 2 Naming for different experimental groups per type of treatment

Treatment →	Absolute	Absolute and non-numerical	Absolute and percentage	Relative and score
Crowdsourced R1—N = 247	Control	R1G1	R1G2	–
Crowdsourced R2—N = 146	Control	–	R2G1	R2G2
Crowdsourced R3—N = 148	Control	–	R3G1	R2G2
Data science and domain experts (Target) N = 54	–	–	TG1	TG2
Domain experts (TargetV)—N = 14	–	–	TVG1	TVG2

old, and with more than 75% of successful completion of other studies in the platform (a parameter suggested by Prolific as a quality-assurance method). Participants were payed a platform-standard rate of 7.5 GBP[6] per hour for participating in the survey. They took an average of 20−25 min to complete the survey. Additionally, we offered a bonus payment of 1 GBP to those who achieved an AUC greater than 0.7. This is common practice and incentivizes conscientious completion of the survey (see, e.g., Green and Chen (2019), Cheng et al. (2019), Yu et al. (2020), Dressel and Farid (2018)).

For the *targeted* studies, participants were recruited through students' mailing lists from two universities in *Catalonia*, as well as social media groups of professionals of data science in countries having the same official language as *Catalonia*. Additionally, we invited professionals from the Justice Department of *Catalonia* to participate; the invitation to participate was done by their Department of Research and Training and the Centre for Legal Studies and Specialised Training from the *Catalonia* regional government.

As a reference, the number of participants in previous crowdsourced user studies is usually a few hundred: 103 in Grgic-Hlaca et al. (2019), 202 in Cheng et al. (2019), 400 in Lin et al. (2020), 462 in Dressel and Farid (2018) and 600 in Green and Chen (2019).

In line with the previous studies, we had 609 participants in total (541 crowdsouced and 68 *targeted*). Nevertheless, we performed a power analysis (independent samples t-test) to test our sample size. Our analysis was made with parameters for a t-test with alpha = 5%, beta = 95% and effect size 0.5, i.e., a "medium" effect size. We obtained a power of 95% with sample size of 88 members for each group (178 in total). These results are in line with the minimum number of participants from previous studies.

[6] Prolific is a UK-based company that uses British pounds as main currency. We follow their advice for average payment per hour.

5 Participants and experimental setup

Along our study, we designed our experiments with different kinds of participants (crowdsourced and targeted with specific knowledge). We use a naming convention to make it space-saving and clear to the reader. Naming are explained at Table 2, where different groups can be compared in terms of experimental setup for a clear understanding.

5.1 Crowdsourced: first round (R1)

In the **first round** (R1) we compared two experimental groups. The **treatment** group was shown the machine prediction and the **control** group was not. In treatment group **G1** machine predictions are shown only as *categorical* information, while in **G2** machine predictions are shown as *categorical and numerical* information. In this round, 247 participants completed the evaluation: 48 in the control group, 100 in treatment group G1, and 99 in treatment group G2. Additionally, 74 participants were excluded, either because they did not complete the survey or did not evaluate all of the eleven cases, or finished the experiment either too fast (less than five minutes) or too slowly (more than one hour).

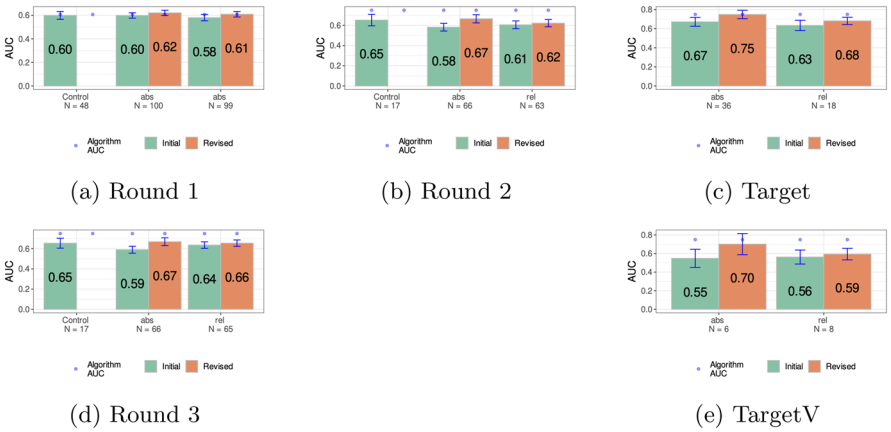
As described in Sect. 4.1.2, we used in R1 a set of cases for which the AUC of the machine predictions was 0.61. To bring this more in line with the observed AUC in the entire dataset (0.76), we exchanged two cases for the second round (R2), and the AUC measured on the new set of cases became 0.75.

5.2 Crowdsourced: second round (R2)

In the **second round** (R2) we compared two experimental groups, where the **treatment** group was shown the machine prediction and the **control** group was not. In treatment group **G1** machine predictions are shown on an *absolute scale* as categorical and numerical information (similar to R1G2), while in **G2** we introduce the machine predictions shown on a *relative scale* as categorical and numerical information. In this round, 146 participants completed the evaluation: 17 in the control group, 66 in treatment group G1, and 63 in treatment group G2. Additionally, 137 participants were excluded for the same reasons as in R1.

5.3 Crowdsourced: third round (R3)

In the **third round** (R3) we compared the same experimental groups like in R2 (G1 and G2) with the purpose of evaluating an iteration of our same interface, but explicitly stating in all screens the fact that they were evaluating violent recidivism (see more in Appendix 3). In this round, 148 participants completed the evaluation: 17 in the control group, 66 in treatment group G1, and 65 in treatment group G2.



(a) Round 1 (b) Round 2 (c) Target

(d) Round 3 (e) TargetV

Fig. 3 Average AUC with 95% confidence interval by group. See Table 7 in Appendix 3 for details

5.4 Targeted study

The *targeted* study seeks to establish the effect (if any) of the participant's background when interacting with the RAI. We used the same experimental setup and treatment groups from crowdsourcing (R2). Due to the limited number of participants, we considered as a baseline the control group of R2 as well as Targeted groups. We used the name *'Targeted'* because we refer to participants with specific professional expertise. We considered both students and professionals with a background either in data science, or in a field relevant to the prison system and the application of RisCanvi, such as psychology, criminology, or social work. For data science, we recruited 14 students at the undergraduate and graduate level, and 11 professionals. For a domain-specific background, we recruited 4 students at the graduate level (Master in Criminology students), and 25 professionals. An additional group of 14 professionals participated (known as *TargetV* group) to contrast the crowdsourced R3 setting. All professionals were recruited trough the Justice Department of *Catalonia* and samples for both targeted groups were part of the same population.

6 Results

In this section, we present our main findings. The main takeaways are:

1. Human predictive accuracy improves after the RAI suggestion.
2. The improvement in the accuracy is also visible over time, after evaluating several cases.
3. Participants disagree on the relative importance of risk factors, this is validated qualitatively in professionals' focus group.
4. Acceptance of automation is limited. All participants foresee automation with, with a clear preference for human discretion.

5. Categorical scales are preferred over numerical scales, and result in higher human predictive accuracy.

6.1 Predictive accuracy

6.1.1 Accuracy

Figure 3 shows the average AUC and corresponding confidence intervals for each experimental group. AUC values depicted in the figure can also be found in Appendix 3, Table 7.

Given the relatively small sample sizes in experimental data, we test the statistical significance of the differences between the experimental groups using a permutation t-test with 999 permutations (see Table 9). For R1 we observe no difference in the *initial* predictions across control and treatment groups, which have AUC from 0.58 to 0.60. However, for R2 we find a significant difference ($p<0.1$) with a higher AUC for the control group (0.65) than for the *initial* prediction of treatment group G1 (0.58) with the absolute scale.

Despite the small number of participants in the *targeted* group, we observe important differences compared to the previous groups. The predictive accuracy of the initial prediction is higher (+0.02 to +0.09 AUC points) than any crowdsourced group. For the *targeted* group G1 (absolute scale) this difference is significant at $p < 0.1$ against R2's G2 and even at $p < 0.05$ against the initial predictions of the other crowdsourced groups (see Appendix 3). Participants from a data science background and domain experts have similar initial AUCs (see Table 7 in Appendix).

We acknowledge a lower AUC on the initial prediction at *TargetV* for both treatment groups. With a reduced number of participants ($N = 14$) we would not consider this difference as important and instead observe this result seem to be inherently noisy in the presence of small samples, hence the large standard deviation observed. For instance, we observe R3 results are within error bars of *TargetV* (see Fig. 3). Besides, the average AUC from *Targeted* and *TargetV* groups together is initially 0.64, and 0.71 revised, which is still higher compared to results obtained in R2 and R3.

The resulting AUC is comparable to previous forensic studies that achieved AUCs on average in the range of 0.65−0.78 using non-algorithmic RAIs (Desmarais et al. 2016; Douglas et al. 2003; Singh et al. 2011).

6.1.2 Self-reported importance of risk items

Having asked to select the top 3 items (risk factors) that participants considered in their risk prediction, we find that crowdsourced and *targeted* participants tend to select the same 10–11 (out of 23) items as more important than the rest. However, among these top 10 items we find that domain experts prefer dynamic factors (i.e., factors that can change), such as *'limited response to psychological treatment'*, while data scientists and crowdsourced participants refer more often than domain experts to static factors (i.e., factors that cannot change), such as *'history of violence'* (details are in Fig. 12 and Table 14 in Appendix 7).

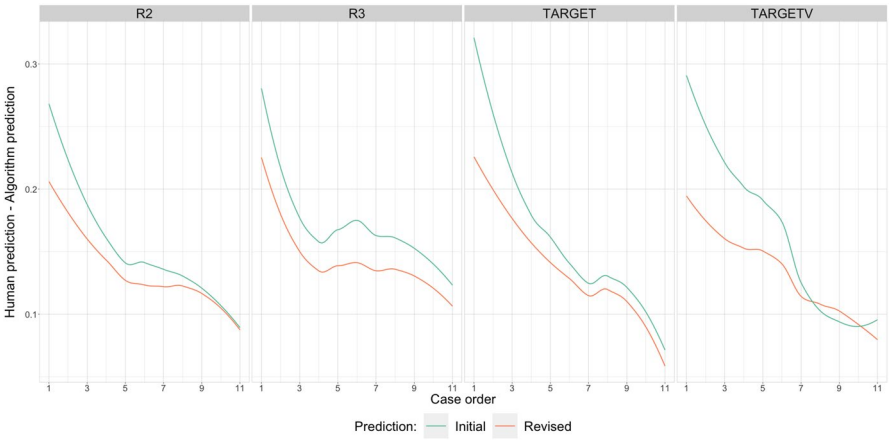


Fig. 4 Average difference between human and algorithm prediction by case order, absolute scale

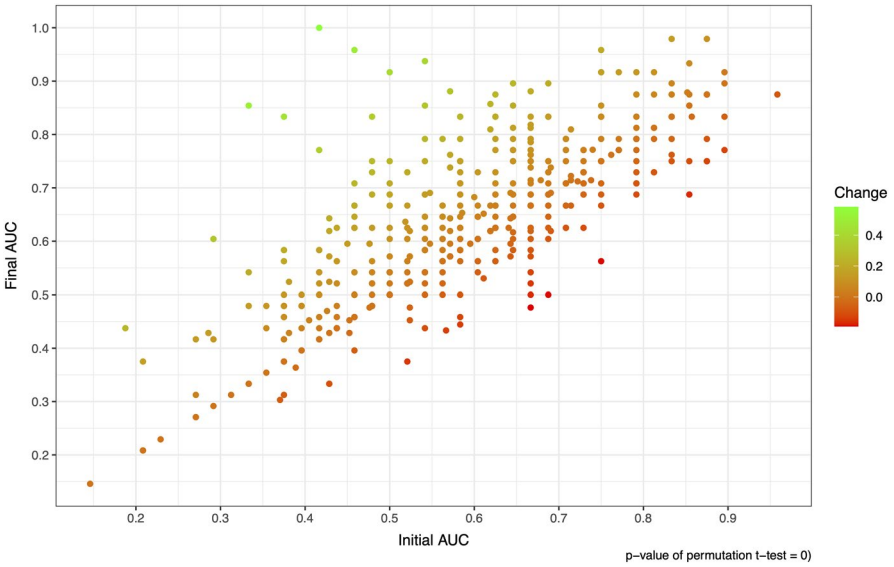


Fig. 5 AUC of participant predictions before and after algorithmic support for participants who received algorithmic support (excludes control group). The p-value of the permutation t-test is $\ll 0.0001$, and the number of permutations for permutation t-test is 999

6.2 Prediction changes due to the RAI

The observed probability of a participant changing a prediction after observing the machine prediction is 20% (19% in G1 and 21% in G2). Crowdsourced participants revised their prediction in about 26% of the cases they examined (27% in

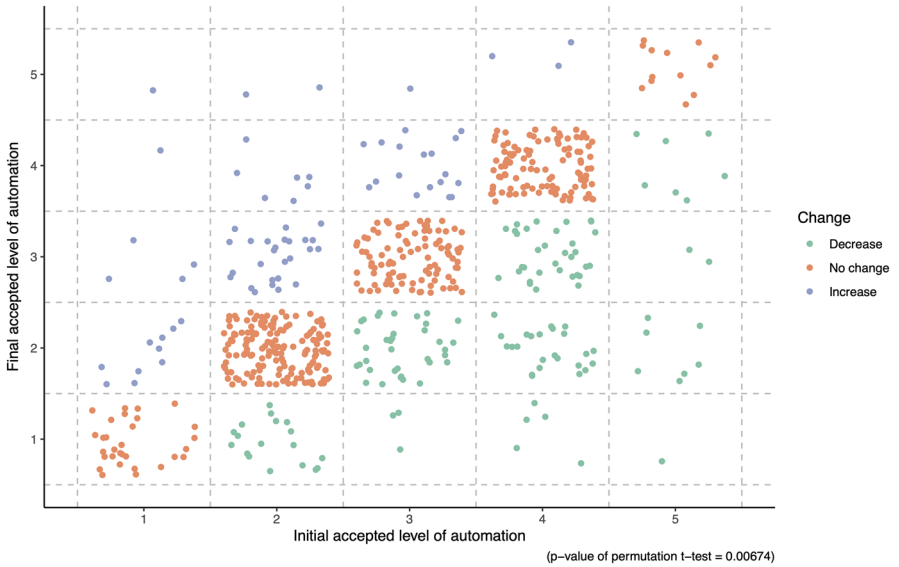


Fig. 6 Distribution of answers about level of automation for participants who received algorithmic support (excludes control group). The p-value of the permutation t-test is < 0.01, and the number of permutations for permutation t-test is 999

G1, 25% in G2). Domain experts revised their prediction in 37% of the cases, and data scientists in only 13% of the cases.

Figure 4 shows the average difference in risk predictions by human and algorithm for each case. *Target* and *TargetV* participants started with predictions that were in general with higher difference than the crowdsourced groups. As they progress through the evaluation tasks, participants tend to align more and more their predictions with the machine predictions (even in their initial predictions) and the difference between initial and revised predictions diminishes. For the last three cases, for the R2 crowdsourced and *Target* groups, which are already close to the machine predictions, do not change, while the R2 and *TargetV* groups maintains a larger difference between initial and revised predictions. We also acknowledge a deviation in the last cases for the TargetV, which might be explained by a lack of attention and could be the cause for the reduced AUC in our results. Nevertheless, this is just an assumption by interpreting the plots and should be contrasted with more evidence.

In general we see that when revising their prediction participants improve their accuracy (Fig 5). By comparing the average AUC in Fig. 3 and Table 7, we can see that revised predictions from crowdsourced groups tend to be more accurate than their initial ones in terms of AUC. This difference is significant for R1's G2 ($p < 0.1$) and for R2's G1 ($p < 0.05$), as shown in Table 9 (in Appendix 3). For the *Target* group, we see an improvement in the range from +0.05 to +0.08 AUC points on average, while for the TargetV the difference is much bigger (+0.03 to +0.15 respectively). In almost all cases, revised predictions by the treatment groups are

more accurate than those of the control groups. However, few of these differences are statistically significant.

In general, the average self-reported confidence is in the range 3.5–3.9 out of 5.0 (1.0 = least confident, 5.0 = most confident), and basically does not change from the initial to the final prediction. The self-reported confidence of crowdworkers is, by a small but statistically significant margin ($p < 0.001$), higher than the one of *targeted* participants (see Appendix 5).

6.3 Preferred level of automation

As shown in Fig. 6, most participants prefer an intermediate level of automation, between levels 2–4 on a scale of 5 levels. While data scientists had an initial level of acceptance with a broader range (levels 1–4), domain experts limited their answers to a more narrow set of choices in intermediate levels of acceptance (levels 2–3). The same figure also shows that most of the treated groups reduce their preferred level of automation after the experiment, meaning they prefer more expert involvement and less reliance on machine predictions.

On average, however, the desired level of acceptance for *targeted* groups concentrated in the middle-low part of the scale: 32% of the data scientists and 48% of the domain experts selected level 3 ('the computational system suggests one option, but the expert can provide an alternative option'). Level 2 ('the computational system suggests some options, but it is the expert who defines the risk level'), was the option selected by 36% of the surveyed data scientists and 38% of the domain experts. Details can be found in Appendix 4, and the description of the automation levels can be found in Appendix 1.

6.4 Risk scales

6.4.1 Categorical versus numerical risk

According to Fig. 3, adding numerical values to the categorical scale does not change the AUC. In G1, where only categorical information is shown, the AUC of revised predictions is slightly higher than the revised predictions in G2, where categorical and numerical values are shown: 0.62 against 0.61 AUC.

6.4.2 Categorical absolute versus relative scales

The results of R2 show that for the initial prediction, the absolute scale (G1) leads to slightly lower AUC compared to the relative scale (G2) (0.58 against 0.61 AUC). However, with algorithmic support, the absolute scale leads to higher AUC than the relative scale (0.67 against 0.62 AUC). Neither of these differences is statistically significant. Additionally, the average AUC of the R2 control group (0.65) is fairly high, and the only higher AUC observation is in the revised predictions using the absolute scale (0.67). The revised and some initial predictions of the *targeted*

participants using the absolute scale significantly outperform all the R1 groups, as well as the R2 groups ($p < 0.05$, see Table 9 in Appendix 3).

6.4.3 Respondent characteristics

With respect to *numeracy*, over 60% of the crowdsourced participants answered correctly 2 or 3 out of the 3 test questions. The *targeted* group had more respondents answering all 3 numeracy questions correctly than crowdworkers, as shown in Table 1: 96% of data scientists obtained results in the highest scores (68% in the top score), while only 59% of domain experts obtained similar results (52% in the top score). We find no correlation between participants' numeracy and their accuracy. The correlation between *decision making style* and *emotional state* with accuracy is not significant either (results are in Appendix 6).

7 Qualitative study

The last study is a qualitative study using focus groups, i.e., groups of participants having a focused discussion on a particular topic (Morgan et al. 1998). The focus groups help us interpret the quantitative results from the *targeted* study, by listening to and learning from participants' experiences and opinions.

7.1 Participants and procedure

Participants (9 women, 4 men) were recruited from the *targeted* experiment, and due to their busy schedules, divided into four groups (FG1–FG4) as follows: **FG1** (N = 3) data scientists; **FG2** (N = 4) domain experts, students from criminology in undergraduate and master levels; **FG3** (N = 2) and **FG4** (N = 4) domain experts working with the Department of Justice, most of them psychologists.

While we did not want to give too much structure to the conversation, to try to uncover new perspectives that we had not thought about, we did prepare a series of questions to stimulate a discussion (available in Appendix 2). The questions address participants' experience with algorithmic predictions and RAIs, their opinion about different scales and categorical/numerical presentation, their understanding of risk factors, and their desired level of automation. Each session lasted between 60 and 90 min and was held online. Following the protocol approved by our Ethics Review Board, participants were asked for their consent to participate and to have the meeting recorded and transcribed. The language of the focus group was the local language of *Catalonia*; the quotes we present in the next section were taken from our transcriptions and paraphrased in English.

7.2 Findings

We focus on our research questions, but note that there were many other insightful comments during the focus groups.

7.2.1 Professional background

All participants were aware that some demographics are over/under represented among prison populations, and thus expected that a RAI trained on such data may lead to discriminatory outcomes. However, the way in which data science participants approached risk prediction was to a large extent based on considering a set of 'anchors' or prototypes (Scurich et al. 2012, p. 13): 'I think about a maximum and a minimum risk. The minimum would be like a person who stole for the first time [...] the maximum would be a killer' (FG1.1). In general, data scientists did not question the presented case characteristics, but domain experts did. Participants in FG3 and FG4 indicated that the risk items, which in RisCanvi only have three levels (Yes/Maybe/No), do not accurately represent the reality of inmates and they were missing the ability to explore or negotiate the risk items during the case evaluations. Furthermore, they indicated that, during the assignment of levels to risk factors, they sometimes 'compensate' higher values in one item with lower values in other items, such that the final risk score matches what they would consider appropriate for the evaluated person. One participant (FG4.1) said that personal biases may also affect the coding of items, as some professionals adopt a more punitive approach, while others take a more protective or rehabilitative approach. Other domain experts agreed with this perspective. Therefore, most professionals expressed the need for teams reviews and validation mechanisms for risk factor codings.

Among domain experts, the psychologists we interviewed were the most concerned about the evidence they collect and the representation of the actual risk. To them, RAIs are tools that add objectivity to their case reports, but their focus was on *how* to present evidence to judges, since these might discard professional reports in favor of the RAI's outcome. Overall, for domain experts RAIs such as *RisCanvi* should be used by a group of experienced evaluators checking one another, and not by one professional alone.

7.2.2 Interpreting numbers

All participants had some training in statistics, and stated that they understand numerical expressions well. Generally, participants preferred a relative scale (e.g., 3.7/10.0) over an absolute scale (e.g., 37%). It is noteworthy how domain experts interpret probabilities.

First, extremely low risks were considered unlikely in practice, since almost everyone can commit a crime at some point.

Second, all interviewed domain experts stated that recidivism risk cannot be eliminated but it could be reduced to an acceptably low level (e.g., reducing the risk from 37% to 20%).

This emphasis on risk reduction is in line with the 'interventions over predictions' debate in the literature (Barabas et al. 2018). Third, domain experts consider a recidivism risk of above 30% as high, and a reason for concern. A risk above 50% was considered difficult -but not impossible- to reduce by treatment/interventions. Overall, domain experts thought of different ranges on the risk spectrum along which inmates

are placed. Data scientists, too, considered different risk ranges, and for some of them even a 50% recidivism risk was not considered 'high'.

7.2.3 Interaction with machine predictions and calibration

Many participants admitted that they went quickly, and without giving it much thought, through the first few evaluations. However, they also noticed that they slowed down to rethink when they felt contested by the algorithm, i.e., when their risk assessment was far from the algorithm's prediction. Data scientists indicated that they reacted to such differences by simply adjusting the risk to half-way between their initial prediction and the one of the algorithm. Domain experts indicated to react similarly in some cases, but they also stressed that they kept their initial prediction when they felt confident about it.

Some of the domain experts believed that they were interacting with exactly the same *RisCanvi* algorithm they use, despite a clear indication in the introduction of the study that this was another algorithm. We believe their experience with the real *RisCanvi* affected their disposition to rely on the machine predictions we presented.

7.2.4 Preferred level of automation

Overall, domain experts and data scientists differed in the level of automation they would prefer, with data scientists being more open to automation. For instance, participant FG1.2 believed that an algorithm could improve enough to make almost-autonomous decisions 'in the future'. This participant considered the errors that could be made by the algorithm were 'acceptable'. In contrast, e.g., FG1.3 was sceptical about using an algorithm for automated decision-making because of the impossibility to solve all algorithm-specific errors.

All participants agreed that algorithmic support is useful in many instances, e.g., to contrast their own predictions, to give them a chance to rethink them, or to provide reassurance about them. Domain experts also considered them useful to train new case workers in writing evaluations. In that regard, participants from FG1 and FG2, expressed that the 'objectivity' of the algorithm could help reduce the effect of the 'emotional' response to the evidence by the professional who is evaluating.

Participants also acknowledged the risk of 'relying too much' on the algorithm, leading to reduced professional responsibility: 'The decision you make is yours, it is yours with your biases and everything, which also brings experience because it sometimes helps you be aware and review your own prejudices' (FG2.1). Another drawback of using a RAI noted by participants was the concern that it may reproduce potentially outdated societal prejudices. To address this concern, domain experts expected frequent updates to the algorithms.

8 Discussion

RQ1: Under which conditions do participants using a RAI to predict recidivism achieve the highest predictive accuracy? Overall, our findings suggest that human decision makers achieve higher accuracy for their risk-assessment when they

are supported by an algorithm. Almost all treatment groups achieve a higher AUC than their corresponding control group after the treatment, although these differences are not statistically significant, particularly in the case of crowdsourced participants (Figs. 3 and 5). Nevertheless, considering the evidence presented in the literature, further studying this phenomenon, possibly with larger populations, is needed. The algorithm also influences human predictions for each decision and over time, as shown in Fig. 4. This further suggests that algorithmic support establishes reference points to human predictions. In Fig. 11 we do not see the influence of algorithmic's accuracy on the improvement of the human decision. Instead, we consider the recurrent use of the tool by professionals as a form of improvement of their own practice. In practical terms, the implementation of *RisCanvi* or any other RAI may have an influence in the long term regardless its accuracy. We consider this should be studied in depth. The lower accuracy of the initial predictions of treatment group participants compared to control group participants is noteworthy. One possible explanation for this is that treated participants put less effort in their initial predictions in anticipation of algorithmic support and a potential opportunity to revise their initial prediction. The exposure to a particular tool is considered important in the field of automated decision-making. Many factors can affect predictive accuracy, as we mention in our limitation Sect. 9.

The finding that *targeted* participants (domain experts and data scientists) outperform crowdsourced participants contradicts the idea from previous work (see Sect. 2) that crowdsourced participants are comparable to domain experts or professionals when testing RAIs. This highlights the importance of testing RAIs in the context of professional knowledge, training and usage.

Finally, using an absolute rather than a relative scale leads to more accurate predictions (in our study in the revised predictions). The focus group further confirmed the preference of professionals for the absolute scale as the one closer to the real application. Our findings agree with Zoe Hilton et al. (2008), who found that risk categories are generally hard to agree upon across professions and individuals, and also with Hanson et al. (2017), who found that categories can be effective following a common agreement in correspondence to ranges of the absolute probability of recidivism. Thus, further studies should focus on the underlying support of numerical information in helping ground categorical distinctions for predictive risk assessment.

RQ2: To what extent do participants rely on the RAI to predict recidivism?

In line with previous studies (e.g., Tan et al. (2018)) humans and algorithms tend to agree on very low and very high risk cases (see Appendix 2, particularly Fig. 7), but there are cases that are difficult to predict for humans, for algorithms, or for both. A promising next step would be to identify cases that are clearly difficult for the machine, and or are potentially difficult to humans. In these cases one could more safely defer to humans, or ask them to invest more time in a specific evaluation, improving efficiency in the design of human-algorithm decision processes. We suggest that any supporting decision system should indicate its confidence for each case, to allow the human to make a more informed decision.

Our findings show that participants prefer a partially automated assistance with a large degree of human discretion. This implies that easy-to-use mechanisms for

overriding or changing the algorithm suggestion are needed, and professionals should be encouraged by their institutions to use them when appropriate. In addition, all experimental groups tend to downgrade the acceptable level of automation after the experiment (see Fig. 6). Explanations for this could be that the differences between human and machine predictions caused the participants to realize strong human oversight was more necessary than what they initially thought.

Finally, the focus group discussions revealed that professionals' reliance in an algorithm could be increased when the algorithm providers ensure good prediction performance and frequent system updates corresponding to new societal and institutional developments. This suggests that *RisCanvi* and possibly other RAIs are elements of negotiation that should be taken with care and without assuming its outcome as objective, and that need frequent updates and audits. So far, all discussions and feedback around the use and the improvement of the algorithm are welcomed by its users. Thus, it is recommended to promote spaces within their organization to hold sessions of discussion and feedback about their experience using a RAI.

9 Limitations and future work

This paper has to be seen in light of some limitations. First, the dataset used for training the algorithm has some drawbacks. It has only about 597 cases, which may affect the algorithm's accuracy; however, we note that its AUC-ROC is in line with that of most recidivism prediction tools. We also note that in this dataset the ground-truth label is *re-incarceration* and not *re-offense*. Re-arrest and re-incarceration are not necessarily a good proxy for re-offense and further exhibits racial and geographical disparities (Fogliato et al. 2021). Since the focus of this study is the assessment of user behaviour (not the algorithm), we do not expect these drawbacks to notably affect our main results. Second, in line with previous work, this study focuses on accuracy as a measure of algorithmic performance. However, decision support algorithms can be evaluated in many different ways (Sambasivan et al. 2021). Third, Fig. 4 shows that participants are still calibrating their predictions after the training phase as they progress through the evaluation tasks, suggesting that the initial training phase may have been too short. The impact should be limited as the majority of the cases are evaluated after this learning curve has flattened.

The generalization of this work to other contexts is restricted by other factors. Due to resource constraints (money to pay crowdsourcing participants, and critically, time availability of domain expert participants), the findings draw from a study centered around 14 cases; a study with more cases would be an improvement, but would be more time-consuming for all participants. These constrains were tackled by selecting a variety of cases that represent different levels of difficulties to assess for humans and the algorithmic system. In addition, as usual in experimental user studies, the crowdsourced participants are not representative of the overall population. Table 1 shows that most have university-level education and good numeracy. Further, we only recruited participants in a single country. Thus, the pool of users might not exhibit a large cultural diversity, a factor that could bias outcomes (Beale and Peter 2008; Lee and Selart 2012). However, we also remark

Table 3 Characteristics of the experimental groups

Type →	Crowdsourced		Crowdsourced			Targeted		
Group ↓	R1	Type	R2	R3	Type	T	TV	Type
Control	48	Abs. scale	17	17	Abs. scale	–	–	–
G1	100	Abs. scale/categorical	66	66	Abs. scale/cat. and num	36	6	Abs. scale/cat. and num
G2	99	Abs. scale/cat. and num	63	65	Rel. scale/cat. and num	18	8	Rel. scale/cat. and num
Total	247		146	148		54	14	

The control groups received no machine predictions

The treatment groups received machine predictions

G1 used an absolute scale indicating a probability (0% to 100%);

G2 used a relative scale indicating a score (0 to 10);

T: Target, TV: TargetV

that crime and recidivism is different in different criminal systems and jurisdictions, and hence RAIs should be evaluated with careful attention to their context (Selbst et al. 2019). With the variations R3 and *TargetV* we tested if explicitly repeating that we refer to violent recidivism in each screen affected the outcome and influence the predictive performance of participants (see Appendix 3) and we have not noticed any substantial change. Nevertheless, we acknowledge that generalizing recidivism definitions can influence in different contexts and these results are not enough representative to reflect these changes. Sample size may be another limitation. While the size of our participant pool in the crowdsourced study (N = 247, N = 146) is in line with previous work, the number of participants in the *targeted* studies (N = 68) is relatively small. We speculated that given that understanding data and probabilities is a complex task, data scientists might be ideal candidates for testing if the results against the crowdsourced and domain expert groups are significantly different. As mentioned in the article a numeracy test didn't fulfill our expectations because criminologists presented a high level as data scientists did. Following this argument, it was confirmed in the focus groups that how data scientists treat information about criminal recidivism is different from professionals and domain experts. We might include impacted or adjacent voices in the study. Despite these limitations in sample size, our results suggest consistent and in some cases statistically significant differences in the outcomes between crowdsourced and *targeted* participants.

It is also important to notice that responses to surveys may incorporate some biases. For example, participants might feel pressure to report socially acceptable answers or suffer from the Hawthorne effect (participants know they are being observed). They also might feel pressured to answer in a short time. However, we required them to answer within a window of 30 min, and most of the participants did it in less than 20 min. All these effects are common when using surveys. Future research is needed to explore the reasons and conditions of these differences. This is particularly important in the public sector, where there is a lack of evidence on how algorithms affect public policies (Zuiderwijk et al. 2021). For example, in the

recidivism prediction context, decision making processes are open, negotiated and mediated, and if a RAI is used for reducing inter-professional communication rather to increase it, it can have adverse effects in decision quality. There is a clear need to pay attention to the usage contexts and the ways in which RAIs are deployed, to reduce the risks of automation and understand better in which conditions the assistance of an algorithm can be most helpful.

Appendix A Additional information about our approach

A.1 Experimental groups

The number of participants in each experimental group is shown in Table 3.

Designing the algorithm

We use logistic regression to predict violent recidivism. The features given as input are the 23 items that determine the REVI score in RisCanvi, plus three demographic features (age, gender, and nationality). The evaluation was done by k-fold cross-validation, i.e., dividing the data into k parts, training on $k-1$ parts and evaluating on the remaining part. The accuracy of the model is 0.76 in terms of AUC-ROC which is the average result over the k runs. Finally, the logistic regression estimates were calibrated, which means that they were transformed to correspond to an estimate of the probability of the outcome.

Datasets

An original dataset with 597 cases was used for creating the algorithm as described above. The dataset is anonymized and shared through a formal collaboration agreement between our university and the Department of Justice of *Catalonia*. This agreement indicates that no personal data is shared with the university.

Semi-synthetic case pool (90 cases)

Although the original dataset did not include personal information, we wanted to make sure that participants never had access to the features of one person. Hence, we created 90 semi-synthetic cases by doing a cross-over of features within a group of similar cases. Each group of cases was selected so that the difference in computed *RisCanvi* between the highest and lowest risk was at most 0.1. The generated case differs by a minimum of one and a maximum of three features from any case in the group, and has a *RisCanvi* risk level within the same risk range of the cases in each group. A preliminary experiment with 31 crowdsourced participants, in which no machine assistance was shown, was used to estimate the difficulty of human risk assessment (i.e., how distant was the prediction from the ground truth) for each case.

Table 4 A confusion matrix of the selection of cases

		Machine prediction	
		Positive	Negative
True re-incarceration	Positive	1	2
	Negative	6	3

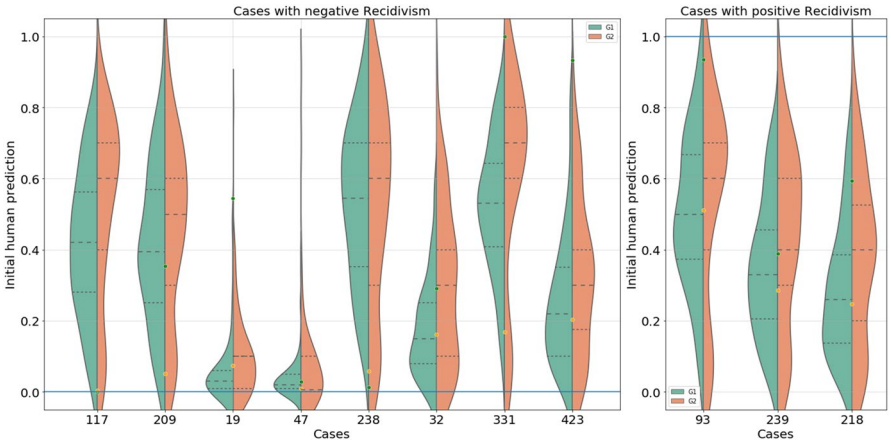


Fig. 7 Comparison of human predictions distribution on 11 experimental cases, on absolute scale setting (G1, in green) and relative scale setting (G2, in orange). The yellow dot indicates the machine prediction for each case. The first 8 cases are non-recidivists (left) and the last 3 are recidivists (right). We remark that all 11 cases were shown in random order to each participant

Case selection (11 cases)

From 90 semi-synthetic cases, 11 cases were selected to be evaluated by participants. This selection was done by sampling 8 non-recidivists and 3 recidivists to have a recidivism rate close to what is observed in *Catalonia*. To perform this sampling, we stratified the cases by human difficulty and machine difficulty into three groups (easy, medium, hard). "Difficulty" means how far, on average, is the prediction from the ground truth. This yields nine classes of difficulty (e.g., "easy" for humans and "hard" for the model) from which we sampled the 11 cases. As we explained in Sect. 4, we exchanged two cases to increase the general AUC-ROC of the entire dataset.

The resulting 11 cases are depicted in Fig. 7, where cases are grouped by ground truth and their risks are predicted by crowdsourced (R2) and targeted studies combined. A confusion matrix in Table 4 shows how cases can be distributed between different categories regarding the machine prediction. It can be noticed that the accuracy of human predictions differs for different cases, and that in some cases, answers are more spread.

Table 5 AUC-ROC in several ML models (sample size = 597)

Model	Logistic regression	XGBoost	Random forest	MLP
AUC-ROC	0.76	0.72	0.73	0.69

Table 6 Feature coefficients in ML-based model (LR) and RisCanvi manually formula by expert

Features	ML-based model (LR)	Original model
Gender	−0.33	−2
Nationality	−0.84	−2
Age	−0.54	Not included
Violent base offense	−0.74	3
Poor childhood adjustment	−0.31	3
Lack of viable future plans	0.42	3
Relevant criminal role	1.25	3
Pro criminal/antisocial attitudes	0.47	3
Intoxication in committing crime	0.09	3
History of violence	0.75	3
Rising crime rates & severity	−0.25	3
Conflict with other inmates	0.40	3
Disciplinary reports	0.82	3
Lack of financial resources	0.31	3
Drug abuse/dependence	0.21	3
Alcohol abuse/dependence	−0.89	3
Limited response to therapy	0.52	3
Low mental ability	−0.16	3
Distance from residence to prison	−0.32	2
Educational level	−0.87	−3
Self-injury attempts/behaviour	0.04	−3
Gender violence victims (women)	−0.08	−3
Recklessness	−0.10	3
Hostility	0.48	3
Family/parental criminal history	0.28	2
Irresponsibility	0.10	2

Details about our algorithm

Feature coeficient comparison with the original algorithm

As mentioned in Sect. 4, our algorithm is based on the same features as *RisCanvi*. In *RisCanvi*, features were weighted by experts in a manual process (using an integer scale from −3 to 3), instead of a machine-learned formula with non-integer coefficients. We note that for this predictive task, besides *RisCanvi* neither the examined literature, nor any deployed systems that we are aware of, use manually crafted formulas with integer

coefficients: they all use statistical and/or machine learning techniques. However, given model multiplicity (multiple models having the same accuracy (Black et al. 2022)), it is not surprising that the manually crafted formula performs relatively well.

We compare in Table 6 the different weights between the manually crafted integer weights and those computed by the logistic regression. Please note that the features in Table 6 are originally coded as in the *RisCanvi* model. The considered values for the features are as follows:

- Gender → 0:male, 1:female
- Nationality → 0:national, 1:foreigner
- Items 17 and 18 → 0:low, 1:high
- Other Items → 0:no, 1:yes

Compared accuracy from other models

During the process of building our algorithms, we tested different models to verify that a logistic regression was the best choice (Table 5).

Surveys

Level of automation survey

This survey was based on the levels of automation proposed by Cummings (Cummings 2004). Cummings proposed ten levels going from 'the computer decides everything and acts autonomously, ignoring the human' (level 10) to 'the computer offers no assistance: a human must take all decisions and actions' (level 1). We reduced the ten levels to five, to make it more understandable and easier to answer for participants.

Question: Would you use a computer system, developed at a university, and based on statistics, to predict the level of risk of violent criminal recidivism?

- Level 1: No, the expert should decide the risk level by himself/herself
- Level 2: Only if the computational system suggests some options, but it is the expert who defines the risk level
- Level 3: Only if the computational system suggests one option, but the expert can provide an alternative option
- Level 4: Only if the computational system suggests one option and the expert can decide to take it or not
- Level 5: Yes, the computational system should decide the risk level by itself

Questions for the focus groups

Questions were used to stimulate the discussion, but we invited participants to comment on any aspect of the experiment.

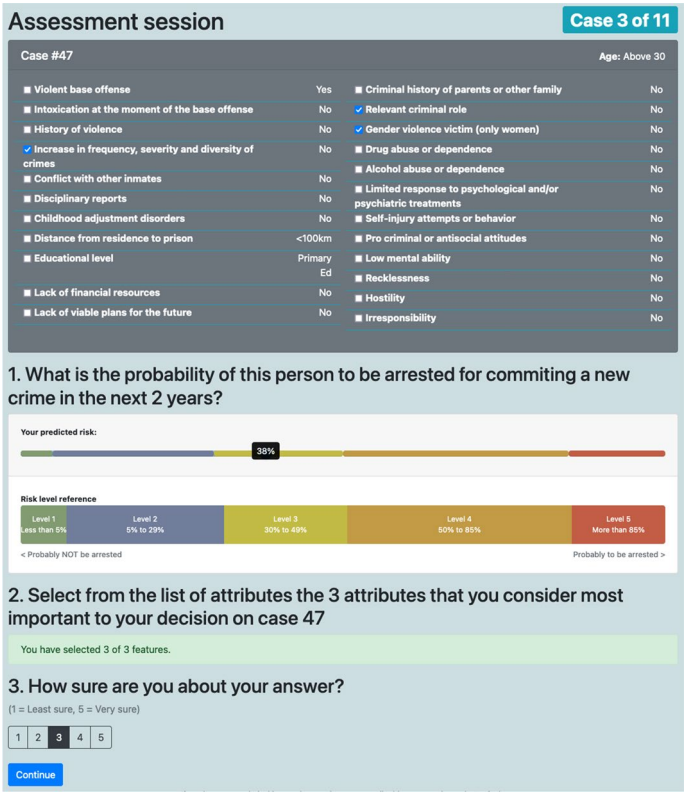


Fig. 8 Evaluation session first page: user prediction and selection of risk factors

- **Q1:** What is your general opinion on Risk Assessment Instruments [explain to participants] in criminal justice settings?
- **Q2:** How do you think the machine prediction in this study works? Could you explain it?
- **Q3:** Do you think that one of these two scales [show them to participants] would be better than the other? Why?
- **Q4:** From the list of case characteristics [show to participants], which were the ones that helped you the most to make a decision about the risk of recidivism?
- **Q5:** Explain, why do you think that these features can help define the prediction of these cases?
- **Q6:** What does a 10% risk mean to you in the context of this study?
- **Q7:** What does a 2 over 10 risk mean for you in the context of the study?
- **Q8:** Despite an improvement in the accuracy, participants tended to rely less on the machine prediction after the experiment, why do you think that it happens? What was your experience?
- **Q9:** Suppose that you can decide to use an algorithm-supported decision making system in this context. What would be the advantages and disadvantages of it?

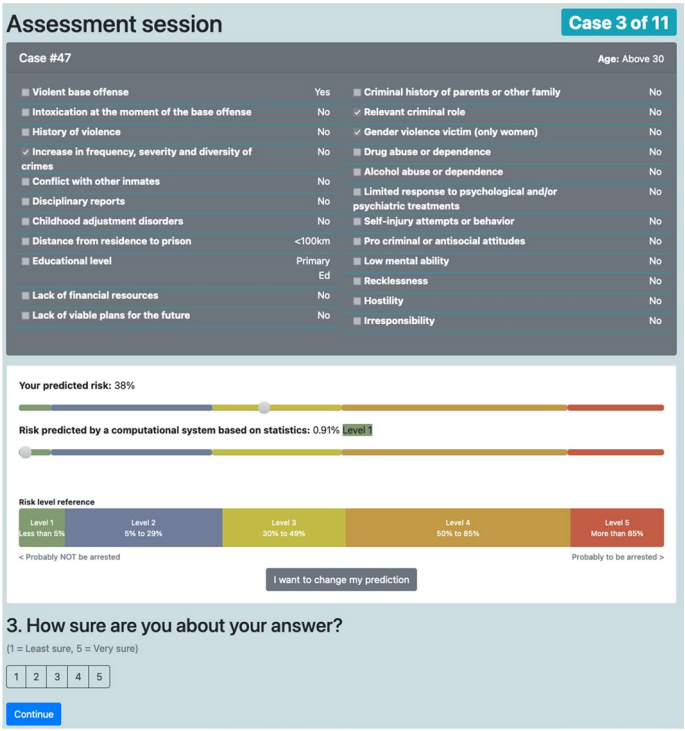


Fig. 9 Evaluation session second page: algorithm prediction and user confirmation

Assessment interface

Figure 8 shows the first page of each evaluation, which is the same for the control and treatment groups. Participants can see each of the risk factors with their corresponding values and select those 3 that they believe are more important to define their prediction. They are asked to define their probability moving the marker on the bar, having as a reference the five levels of risk that depend on the type of scale used. Before moving to the next page they have to assign a value to their confidence in their answer. Figure 9 shows the second page with the machine assistance, which only the treatment groups see. It shows the algorithm prediction in a similar bar/scale, compared to the participant's prediction. Then, the participant has the possibility to change their own prediction and provide a confidence score.

Figure 10 shows the interfaces tested in the crowdsourced R1 study. The control group only sees their own prediction without any feedback. The G1 (bottom) is able to see their prediction compared with the algorithm's prediction, while G2 (middle) is able to see also the most important features that defines the algorithm's prediction (explainers).

For the R3 and TargetV groups we speculated if remarking that participants were evaluating violence recidivism in the first screen and all case evaluation pages it would give a different outcome. For example, for the G1 (absolute scale), instead of

Fig. 10 Alternative treatment groups in crowdsourced R1 study

asking 'What is the probability of this person to be arrested for committing a new crime in the next 2 years?', we changed to 'What is the probability of this person of being incarcerated for committing a new violent crime in the next two years?'. In the first screen we also specified the differences between violent recidivism and general recidivism. We also exchanged the term re-arrest for re-incarceration for the possible effects on interpretation.

Appendix C Additional results

C.1 Accuracy results per sub-group

In Table 7 we present the results from different groups, including targeted groups of data scientists and domain experts, separating students and professionals. In the table we also include the Brier Score (lower is better), which is consistent with the AUC-ROC results (higher is better).

Table 7 AUC-ROC (and Brier score in **gray**) before and after algorithmic support by experimental group, type of risk scale and numerical expression

Groups →	Control	R1G1		R1G2 & R2G1 & TG1		R2G2 & TG2	
Treatment →	Absolute	Absolute & non-numerical		Absolute & percentage		Relative & score	
Participant type ↓		Before	After	Before	After	Before	After
Crowdsourced R1—N = 247	0.599 (**0.277**)	0.598 (**0.273**)	0.621 (**0.271**)	0.576 (**0.291**)	0.607 (**0.280**)	–	–
Crowdsourced R2—N = 146	0.652 (**0.262**)	–	–	0.582 (**0.265**)	0.665 (**0.240**)	0.606 (**0.271**)	0.622 (**0.269**)
Crowdsourced R3—N = 148	0.654 (**0.261**)	–	–	0.590 (**0.263**)	0.669 (**0.239**)	0.636 (**0.270**)	0.655 (**0.277**)
Data science (students)—N = 14	–	–	–	0.652 (**0.229**)	0.690 (**0.203**)	0.589 (**0.269**)	0.667 (**0.266**)
Data science (professionals)—N = 11	–	–	–	0.736 (**0.219**)	0.729 (**0.218**)	0.667 (**0.269**)	0.719 (**0.267**)
Domain experts (students)—N = 4	–	–	–	0.771 (**0.211**)	0.833 (**0.169**)	0.750 (**0.264**)	0.729 (**0.264**)
Domain experts (professionals T)—N = 25	–	–	–	0.636 (**0.236**)	0.773 (**0.191**)	0.637 (**0.270**)	0.673 (**0.267**)
Domain experts (professionals TV)—N = 14	–	–	–	0.549 (**0.270**)	0.701 (**0.221**)	0.562 (**0.272**)	0.594 (**0.270**)

Table 8 Aditional metrics: F1-Score (weighted) [F1], True Positive Rate [T] and False Positive Rate [F] before and after algorithmic support by experimental group, type of risk scale and numerical expression

Groups →	R1G1		R1G2 & R2G1 & TG1		R2G2 & TG2	
Treatment →	Absolute & non-numerical		Absolute & percentage		Relative & score	
Participant type ↓	Before	After	Before	After	Before	After
Crowdsourced R1—N = 247	F1:0.53 T:0.86 F:0.64	F1:0.61 T:0.28 F:0.16	F1:0.54 T: 0.89 F: 0.63	F1:0.57 T:0.24 F:0.19	–	–
Crowdsourced R2—N = 146	–	–	F1:0.48 T:0.93 F:0.73	F1:0.65 T:0.24 F:0.18	F1:0.52 T:0.83 F:0.61	F1:0.61 T:0.20 F:0.22
Crowdsourced R3—N = 148	–	–	F1:0.44 T:0.93 F:0.73	F1: 0.62 T:0.24 F:0.23	F1:0.52 T:0.81 F:0.60	F1:0.60 T:0.19 F:0.22
Domain experts (professionals T)—N = 25	–	–	F1:0.58 T:0.74 F:0.51	F1: 0.59 T:0.18 F:0.22	F1:0.60 T:0.65 F:0.46	F1:0.60 T:0.19 F:0.20
Domain experts (professionals TV)—N = 14	–	–	F1:0.58 T:0.72 F:0.50	F1: 0.62 T:0.33 F:0.27	F1:0.45 T:0.83 F:0.69	F1:0.60 T:0.21 F:0.23

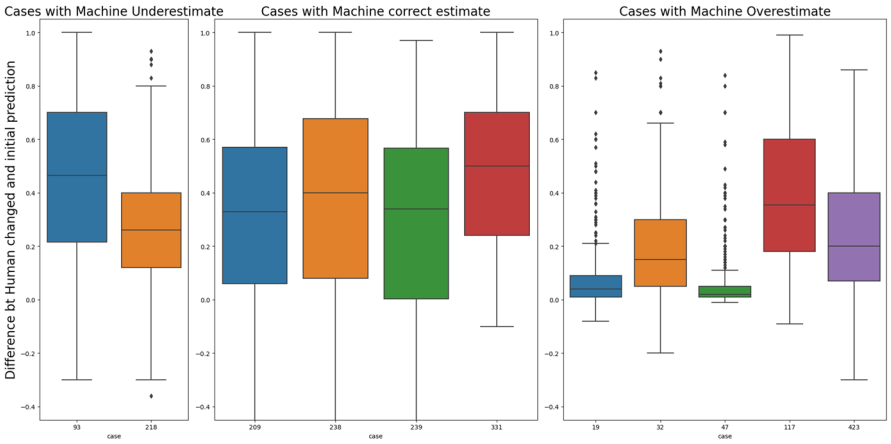


Fig. 11 Difference between the initial and changed prediction shown in three buckets (underestimate, correct estimate within a given tolerance, overestimate) for each of the cases

In Table 8 we include additional metrics for specific groups. We choose to include only the crowdsourced and the professional groups from our study since those are more relevant to our research questions. F1 score was calculated by defining a threshold that yields in the maximum resuld from the precision-recall curve. The result shows that F1 score improves in both scales after the influence of the RAI.

In addition, in Table 9 we show results from a t-test to show differences between crowdsourced and targeted groups.

From the analyzed cases, we also computed the revised prediction differences after observing the RAI. In Fig. 11 we split the cases according to the RAI's performance at estimating its risk (underestimate, correct, or overestimate) and plot the difference between the initial and changed prediction from the participants in all the cases. The effect of RAI performance is smaller than the effect of number in the sequence of cases that a participant sees, as shown in Table 4.

Preferred level of automation and experience

In addition to the level of automation question, we asked participants about their previous experience with Risk Assessment Instruments (RAI) on a scale from 1 (This is the first time I heard or read about risk assessment instruments) to 5 (I have used this kind of tools more than one time). In Table 10 we can see that targeted participants report more previous experience than crowdsourced participants. We can also see that the preferred level of automation is in general lower at the end of the experiment than at the beginning, in all cases except for the control groups.

Table 9 Permutation t-tests with 999 permutations for differences between AUC-ROC of subgroups, showing only differences that are significant at $p < 0.1$. In most cases the relation is that crowdsourced groups are less accurate than targeted groups, and predictions before machine assistance are less accurate than predictions after machine assistance. Cases where this relationship is inverted are more rare. The relationship signs (<, >) indicate which group has a higher AUC

Group 1	Relation	Group 2	p-value
Crowd R2 Control Before	>	Crowd R2 G1 Before	0.089
Crowd R2 Control After	>	Crowd R2 G1 Before	0.092
Crowd R2 Control Before	<	Target G1 After	0.016
Crowd R2 Control After	<	Target G1 After	0.013
Crowd R2 Control Before	>	TargetV G1 Before	0.08
Crowd R2 Control After	>	TargetV G1 Before	0.094
Crowd R2 Control Before	>	TargetV G2 Before	0.088
Crowd R2 Control After	>	TargetV G2 Before	0.094
Crowd R2 G1 Before	>	Crowd R2 Control Before	0.085
Crowd R2 G1 Before	>	Crowd R2 Control After	0.096
Crowd R2 G1 Before	<	Crowd R2 G1 After	0.004
Crowd R2 G1 After	<	Crowd R2 G1 Before	0.003
Crowd R2 G1 After	<	Crowd R2 G2 Before	0.04
Crowd R2 G1 Before	>	Crowd R2 Control Before	0.08
Crowd R2 G1 Before	>	Crowd R2 Control After	0.081
Crowd R2 G1 Before	<	Crowd R2 G1 After	0.001
Crowd R2 G1 After	<	Crowd R2 G1 Before	0.008
Crowd R2 G1 Before	<	Crowd R3 G2 Before	0.036
Crowd R2 G1 Before	<	Crowd R3 G2 After	0.005
Crowd R2 G1 Before	<	Target G1 Before	0.006
Crowd R2 G1 Before	<	Target G1 After	0.0
Crowd R2 G1 After	<	Target G1 After	0.008
Crowd R2 G1 Before	<	Target G2 After	0.011
Crowd R2 G1 Before	>	TargetV G1 After	0.079
Crowd R2 G1 After	>	TargetV G2 Before	0.097
Crowd R2 G2 Before	<	Crowd R2 G1 After	0.041
Crowd R2 G2 Before	<	Crowd R2 G1 After	0.025
Crowd R2 G2 After	>	Crowd R2 G1 After	0.09
Crowd R2 G2 Before	>	Crowd R3 G2 After	0.054
Crowd R2 G2 Before	<	Target G1 Before	0.037
Crowd R2 G2 Before	<	Target G1 After	0.0
Crowd R2 G2 After	>	Target G1 Before	0.099
Crowd R2 G2 After	<	Target G1 After	0.0
Crowd R2 G2 Before	>	Target G2 After	0.051
Crowd R2 G2 After	>	Target G2 After	0.098
Crowd R2 Control Before	>	Crowd R2 G1 Before	0.078
Crowd R2 Control After	>	Crowd R2 G1 Before	0.078
Crowd R2 Control Before	>	Crowd R2 G1 Before	0.084
Crowd R2 Control After	>	Crowd R2 G1 Before	0.091

Table 9 (continued)

Group 1	Relation	Group 2	p-value
Crowd R2 Control Before	<	Target G1 After	0.012
Crowd R2 Control After	<	Target G1 After	0.014
Crowd R2 Control Before	>	TargetV G1 Before	0.05
Crowd R2 Control After	>	TargetV G1 Before	0.051
Crowd R2 Control Before	>	TargetV G2 Before	0.056
Crowd R2 Control After	>	TargetV G2 Before	0.053
Crowd R2 G1 Before	>	Crowd R2 Control Before	0.099
Crowd R2 G1 Before	<	Crowd R2 G1 After	0.007
Crowd R2 G1 After	<	Crowd R2 G1 Before	0.002
Crowd R2 G1 After	<	Crowd R2 G2 Before	0.023
Crowd R2 G1 After	>	Crowd R2 G2 After	0.083
Crowd R2 G1 Before	>	Crowd R2 Control Before	0.091
Crowd R2 G1 Before	>	Crowd R2 Control After	0.085
Crowd R2 G1 Before	<	Crowd R2 G1 After	0.003
Crowd R2 G1 After	<	Crowd R2 G1 Before	0.004
Crowd R2 G1 Before	>	Crowd R3 G2 Before	0.056
Crowd R2 G1 Before	<	Crowd R3 G2 After	0.008
Crowd R2 G1 Before	<	Target G1 Before	0.005
Crowd R2 G1 Before	<	Target G1 After	0.0
Crowd R2 G1 After	<	Target G1 After	0.014
Crowd R2 G1 Before	<	Target G2 After	0.01
Crowd R2 G1 Before	>	TargetV G1 After	0.072
Crowd R2 G1 After	>	TargetV G1 Before	0.079
Crowd R2 G1 After	>	TargetV G2 Before	0.073
Crowd R3 G2 Before	<	Crowd R2 G1 Before	0.036
Crowd R3 G2 After	<	Crowd R2 G1 Before	0.004
Crowd R3 G2 After	<	Crowd R2 G2 Before	0.049
Crowd R3 G2 Before	>	Crowd R2 G1 Before	0.06
Crowd R3 G2 After	<	Crowd R2 G1 Before	0.007
Group 1	Relation	Group 2	p-value
Crowd R3 G2 Before	<	Target G1 After	0.0
Crowd R3 G2 After	<	Target G1 After	0.001
Crowd R3 G2 After	>	TargetV G1 Before	0.057
Crowd R3 G2 After	>	TargetV G2 Before	0.057
Target G1 After	<	Crowd R2 Control Before	0.014
Target G1 After	<	Crowd R2 Control After	0.013
Target G1 Before	<	Crowd R2 G1 Before	0.005
Target G1 After	<	Crowd R2 G1 Before	0.0
Target G1 After	<	Crowd R2 G1 After	0.012
Target G1 Before	<	Crowd R2 G2 Before	0.044
Target G1 After	<	Crowd R2 G2 Before	0.0
Target G1 After	<	Crowd R2 G2 After	0.0

Table 9 (continued)

Group 1	Relation	Group 2	p-value
Target G1 After	<	Crowd R2 Control Before	0.013
Target G1 After	<	Crowd R2 Control After	0.013
Target G1 Before	<	Crowd R2 G1 Before	0.008
Target G1 After	<	Crowd R2 G1 Before	0.0
Target G1 After	<	Crowd R2 G1 After	0.013
Target G1 After	<	Crowd R3 G2 Before	0.0
Target G1 After	<	Crowd R3 G2 After	0.001
Target G1 Before	<	Target G1 After	0.019
Target G1 After	<	Target G1 Before	0.019
Target G1 After	<	Target G2 Before	0.004
Target G1 After	>	Target G2 After	0.057
Target G1 Before	>	TargetV G1 Before	0.052
Target G1 After	<	TargetV G1 Before	0.002
Target G1 Before	<	TargetV G2 Before	0.048
Target G1 After	<	TargetV G2 Before	0.001
Target G1 After	<	TargetV G2 After	0.004
Target G2 After	<	Crowd R2 G1 Before	0.012
Target G2 After	>	Crowd R2 G2 Before	0.054
Target G2 After	>	Crowd R2 G2 After	0.098
Target G2 After	<	Crowd R2 G1 Before	0.012
Target G2 Before	<	Target G1 After	0.003
Target G2 After	>	Target G1 After	0.058
Target G2 After	<	TargetV G1 Before	0.006
Target G2 After	<	TargetV G2 Before	0.003
Target G2 After	<	TargetV G2 After	0.027
TargetV G1 Before	>	Crowd R2 Control Before	0.09
TargetV G1 Before	>	Crowd R2 Control After	0.091
TargetV G1 After	>	Crowd R2 G1 Before	0.079
TargetV G1 Before	>	Crowd R2 Control Before	0.05
TargetV G1 Before	>	Crowd R2 Control After	0.051
TargetV G1 Before	>	Crowd R2 G1 After	0.086
TargetV G1 After	>	Crowd R2 G1 Before	0.068
TargetV G1 Before	>	Crowd R3 G2 After	0.056
TargetV G1 Before	>	Target G1 Before	0.052
TargetV G1 Before	<	Target G1 After	0.001
TargetV G1 Before	<	Target G2 After	0.005
TargetV G1 Before	>	TargetV G1 After	0.076
TargetV G1 After	>	TargetV G1 Before	0.076
TargetV G1 After	>	TargetV G2 Before	0.062
TargetV G2 Before	>	Crowd R2 Control Before	0.087
TargetV G2 Before	>	Crowd R2 Control After	0.099
TargetV G2 Before	>	Crowd R2 G1 After	0.099

Table 9 (continued)

Group 1	Relation	Group 2	p-value
TargetV G2 Before	<	Crowd R2 Control Before	0.049
TargetV G2 Before	>	Crowd R2 Control After	0.052
TargetV G2 Before	>	Crowd R2 G1 After	0.069
TargetV G2 Before	>	Crowd R3 G2 After	0.059
TargetV G2 Before	>	Target G1 Before	0.054
TargetV G2 Before	<	Target G1 After	0.0
TargetV G2 After	<	Target G1 After	0.003
TargetV G2 Before	<	Target G2 After	0.003
TargetV G2 After	<	Target G2 After	0.024
TargetV G2 Before	>	TargetV G1 After	0.061

Table 10 Users' experience with RAIs (scale of 1–5) and preferred level of automation (scale of 1–5, as described in Sect. 1) at the start and at the end of the study, including standard deviation values

Experimental groups	Experience (1 = least, 5 = most)	Preferred level of automation (1 = no automation, 5 = fully automated)	
		Start	End
Crowd R1—Control	1.62 ± 0.76	2.88 ± 1.02	2.96 ± 1.03
Crowd R1—G1	1.79 ± 0.99	3.08 ± 1.07	3.01 ± 1.05
Crowd R1—G2	1.70 ± 0.82	3.01 ± 1.08	2.87 ± 1.10
Crowd R2—Control	1.29 ± 0.47	2.82 ± 1.07	2.59 ± 1.00
Crowd R2—G1	1.48 ± 0.67	2.61 ± 1.03	2.48 ± 1.06
Crowd R2—G2	1.70 ± 0.85	3.00 ± 1.03	2.60 ± 0.91
Crowd R3—Control	2.12 ± 1.17	2.94 ± 0.83	2.88 ± 0.86
Crowd R3—G1	1.65 ± 0.81	2.55 ± 1.12	2.30 ± 0.93
Crowd R3—G2	1.75 ± 0.90	2.53 ± 0.99	2.60 ± 0.97
Target G1	3.02 ± 1.43	2.75 ± 0.90	2.68 ± 0.84
Target G2	3.16 ± 1.44	3.11 ± 0.83	3.05 ± 0.87
TargetV G1	5.00 ± 0.00	3.00 ± 1.26	3.33 ± 1.03
TargetV G2	4.50 ± 1.41	3.00 ± 0.76	3.50 ± 0.76

Self-reported confidence

We observe that self-reported confidence is stable across all subgroups. For each case evaluation, participants had to answer their level of confidence on a likert scale (1–5). After seeing the algorithm prediction and with the opportunity to change their prediction or not, they are asked about their confidence level again (Tables 11 and 12). Results are shown in Table 11.

Table 11 Average self-reported confidence by subgroup, with standard deviation values, before and after seeing the machine prediction. 1 = least confident, 5 = most confident

Study	Group	Before	After
Crowd (R2)	G1	3.89 ± 0.89	3.82 ± 0.98
Crowd (R2)	G2	3.90 ± 0.77	3.82 ± 0.95
Crowd (R3)	G1	3.80 ± 0.58	3.78 ± 0.64
Crowd (R3)	G2	3.75 ± 0.58	3.69 ± 0.65
Targeted	G1	3.48 ± 0.71	3.45 ± 0.85
Targeted	G2	3.57 ± 0.81	3.52 ± 0.97
Target	Domain experts	3.47 ± 0.73	3.52 ± 0.82
Target	Data scientists	3.55 ± 0.76	3.43 ± 1.04
TargetV	Domain experts	3.59 ± 0.49	3.60 ± 0.50

Table 12 Pairwise permutation t-test for self-reported confidence with 999 permutations, only p <0.001 is shown

Group 1	Group 2	p-value
Round2 G1 Initial	Target G1 Initial	<0.001****
Round2 G1 Final	Target G1 Initial	<0.001****
Round2 G1 Initial	Target G2 Initial	<0.001****
Round2 G1 Final	Target G2 Initial	<0.001****
Round2 G2 Initial	Round3 G2 Initial	<0.001****
Round2 G2 Initial	Target G1 Initial	<0.001****
Round2 G2 Initial	Target G2 Initial	<0.001****
Round2 G2 Initial	TargetV G2 Final	<0.001****
Round3 G1 Initial	Target G1 Initial	<0.001****
Round3 G1 Initial	Target G2 Initial	<0.001****
Round3 G2 Initial	Round2 G2 Initial	<0.001****
Round3 G2 Initial	Target G1 Initial	<0.001****
Target G1 Initial	Round2 G1 Initial	<0.001****
Target G1 Initial	Round2 G1 Final	<0.001****
Target G1 Initial	Round2 G2 Initial	<0.001****
Target G1 Initial	Round3 G1 Initial	<0.001****
Target G1 Initial	Round3 G2 Initial	<0.001****
Target G2 Initial	Round2 G1 Initial	<0.001****
Target G2 Initial	Round2 G1 Final	<0.001****
Target G2 Initial	Round2 G2 Initial	<0.001****
Target G2 Initial	Round3 G1 Initial	<0.001****
TargetV G1 Initial	Round2 G2 Initial	<0.001****
TargetV G2 Final	Round2 G2 Initial	<0.001****

Decision-making style and emotional state results

During the survey we included two surveys to measure the current emotional state (e.g., joyful, sad, anger) and decision making style (e.g., rational, intuitive).

Table 13 DGMS and VAS results

Survey	Crowd. (R1)			Crowd. (R2)			Crowd. (R3)			Target		TargetV	
	GC	G1	G2	GC	G1	G2	GC	G1	G2	G1	G2	G1	G2
VAS													
Joy	3.27	3.51	3.45	3.59	3.55	3.92	3.47	3.59	3.46	3.89	3.72	3.59	3.460
Sad	2.29	2.08	2.24	2.29	2.17	1.63	2.18	2.20	2.02	1.97	2.11	2.20	2.020
Angry	1.60	1.65	1.59	1.71	1.83	1.46	1.76	1.56	1.58	1.56	1.22	1.56	1.58
Surprise	1.75	1.77	1.86	1.71	1.86	2.13	1.59	1.89	1.94	1.72	2.06	1.89	1.94
Relax	3.79	3.59	3.74	3.47	3.55	3.94	3.82	3.50	3.33	3.750	3.50	3.50	3.750
Energy	2.94	3.05	2.80	2.88	3.09	3.32	2.94	3.11	3.31	2.91	3.00	3.11	2.91
GDMS													
Rational	0.61	0.61	0.64	0.63	0.60	0.62	0.60	0.60	0.62	0.58	0.59	0.49	0.580
Intuitive	0.72	0.72	0.71	0.72	0.75	0.74	0.69	0.71	0.71	0.71	0.68	0.67	0.61
Dependent	0.59	0.59	0.64	0.55	0.58	0.59	0.59	0.60	0.60	0.57	0.53	0.51	0.52
Avoidant	0.59	0.59	0.64	0.61	0.59	0.61	0.56	0.60	0.60	0.57	0.54	0.42	0.455
Spontaneous	0.68	0.69	0.71	0.68	0.69	0.71	0.67	0.73	0.68	0.71	0.70	0.67	0.67

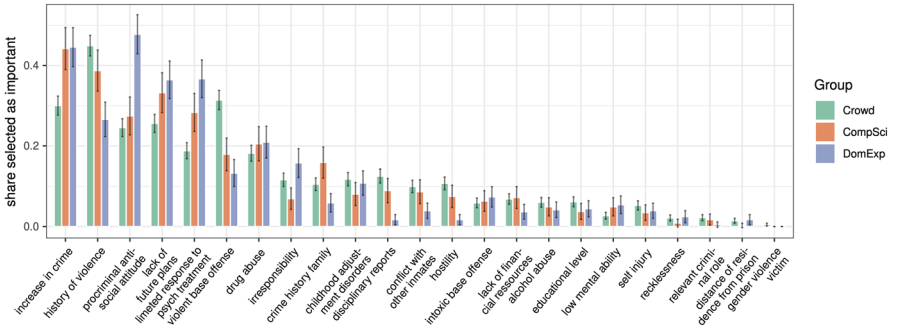


Fig. 12 Amount of times that each item was selected as important to the decision, by participants with different backgrounds (crowdsourced, data science and domain experts)

We used VAS (Portela and Granell-canut 2017) for the emotional state, and GDMS (Scott and Bruce 1995) for the decision making style.

Results reflected in Table 13 are similar across subgroups. Common emotional states reported are joyful, relaxed, and energized. In general, intuitive and spontaneous decision making appears with higher levels than rational.

RisCanvi items considered as most important

As explained in Sect. 6.1.2, the top items (features) selected as important for most participants tend to be the same (Fig. 12). Targeted groups of data scientists and domain experts selected the same top five items, albeit in a different ordering, and

Table 14 Top 5 items (features) listed as most important for making decisions by different background

	Crowdsourced	Targeted: data science	Targeted: domain experts
1st	History of violence	Increase in frequency, severity and diversity of crimes	Pro criminal or antisocial attitude
2nd	Violent base offense	History of violence	Increase in frequency, severity and diversity of crimes
3rd	Increase in frequency, severity and diversity of crimes	Lack of viable plans for the future	Limited response to psychological/psychiatric treatment
4th	Lack of viable plans for the future	Limited response to psychological/psychiatric treatment	Lack of viable plans for the future
5th	Pro criminal or antisocial attitude	Pro criminal or antisocial attitude	History of violence

their top items overlap to some extent with those of crowdsourced participants (Table 14).

Acknowledgements We thanks the collaboration from the Justice Department of Catalonia. We thanks Emilia Gomez from the Joint Research Centre, European Commission; Alexandra Chouldechova and Riccardo Fogliato from Carnegie Mellon University, for their invaluable contributions and support.

Author Contributions The manuscript was coordinated by Manuel Portela and Carlos Castillo, with important contributions from Songül Tolan and Marzieh Karimi-Haghighi on all sections. It was edited by the four authors equally. Antonio Andres Pueyo contributed with his expertise on the domain, providing useful feedback during the process. All authors reviewed the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work has been partially supported by the HUMAINT programme (Human Behaviour and Machine Intelligence), Centre for Advanced Studies, Joint Research Centre, European Commission through the Expect contract CT-EX2019D347180. Additionally, it received the support from the EU-funded project "SoBigData++" (grant agreement 871042).

Data Availability Interaction data and selected cases used can be provided upon request.

Declarations

Conflict of interest There is no conflict of interest or competing interests from any of the authors.

Ethics approval The research project was approved by the University's ethical committee (CIREP) certifying that complies with the data protection legal framework, namely, with European General Data Protection Regulation (EU) 2016/679 -GDPR- and Spanish Organic Law 3/2018, of December 5th, on Protection of Personal Data and digital rights guarantee -LOPDGDD-.

Consent to participate All participants were informed and gave their consent to participate in the study.

Consent for publication All participants gave their consent to publish the data collected in the surveys and related to their participation in the study.

References

Andrés-Pueyo A, Arbach-Lucioni K, Redondo S, Kroner J, Stephen Wormith SL, Desmarais Z (2018) The riscanvi: a new tool for assessing risk for violence in prison and recidivism. In: Recidivism risk assessment: a handbook for practitioners. Wiley, pp. 255–268

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias: there's software used across the country to predict future criminals and it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing

Bansak K (2019) Can nonexperts really emulate statistical learning methods? A comment on "the accuracy, fairness, and limits of predicting recidivism". Polit Anal 370–380

Bao M *et al.* (2021) It's compaslicated: the messy relationship between rai datasets and algorithmic fairness benchmarks. arXiv preprint. arXiv:2106.05498
Barabas C, Virza M, Dinakar K, Ito J, Zittrain J (2018) Interventions over predictions: reframing the ethical debate for actuarial risk assessment, PMLR, pp 62–76
Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671
Batastini AB et al (2019) Does the format of the message affect what is heard? a two-part study on the communication of violence risk assessment data. J Forensic Psychol Res Pract 19:44–71. https://doi.org/10.1080/24732850.2018.1538474
Beale R, Peter C (2008) The role of affect and emotion in HCI. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 4868 LNCS, pp 1–11
Berk R (2017) An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. J Exp Criminol 13:193–216
Binns R, Veale M (2021) Is that your final decision? Multi-stage profiling, selective effects, and article 22 of the GDPR. Int Data Privacy Law 00:1–14
Black E, Raghavan M, Barocas S (2022) Model multiplicity: opportunities, concerns, and solutions, FAccT'22. Association for Computing Machinery, New York, pp 850–863. https://doi.org/10.1145/3531146.3533149
Burton JW, Stein M-K, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Mak 33:220–239
Chancey ET, Bliss JP, Yamani Y, Handley HA (2017) Trust and the compliance-reliance paradigm: the effects of risk, error bias, and reliability on trust and dependence. Hum Factors 59:333–345
Cheng H-F et al (2019) Explaining decision-making algorithms through UI. ACM Press, New York, pp 1–12. http://dl.acm.org/citation.cfm?doid=3290605.3300789
Chiusi F, Fischer S, Kayser-Bril N, Spielkamp M (2020) Automating society report 2020. Tech Rep, AlgorithmWatch . https://automatingsociety.algorithmwatch.org
Cummings ML (2004) Automation bias in intelligent time critical decision support systems. In: Collection of technical papers—AIAA 1st intelligent systems technical conference, vol 2, pp 557–562
Dahle K-P, Biedermann J, Lehmann RJ, Gallasch-Nemitz F (2014) The development of the crime scene behavior risk measure for sexual offense recidivism. Law Hum Behav 38:569
De-Arteaga M, Fogliato R, Chouldechova A (2020) A case for humans-in-the-loop: decisions in the presence of erroneous algorithmic scores, pp 1–12. ACM, New York. https://doi.org/10.1145/3313831.3376638. arXiv:2002.08035
Desmarais S, Singh J (2013) Risk assessment instruments validated and implemented in correctional settings in the united states. Council of State Governments, Lexington, KY
Desmarais SL, Johnson KL, Singh JP (2016) Performance of recidivism risk assessment instruments in us correctional settings. Psychol Serv 13:206
Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144:114
Douglas KS, Ogloff JR, Hart SD (2003) Evaluation of a model of violence risk assessment among forensic psychiatric patients. Psychiatr Serv 54:1372–1379
Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4:1–6
Du N, Huang KY, Yang XJ (2019) Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. Hum Factors
Fogliato R, Chouldechova A, Lipton Z (2021) The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. arXiv preprint. arXiv:2109.01443
Fogliato R, Xiang A, Lipton Z, Nagin D, Chouldechova A (2021) On the validity of arrest as a proxy for offense: race and the likelihood of arrest for violent crimes. arXiv preprint. arXiv:2105.04953
Goel S, Shroff R, Skeem JL, Slobogin C (2019) The accuracy, equity, and jurisprudence of criminal risk assessment. *SSRN Electr J* 1–21
Green B (2020) The false promise of risk assessments: epistemic reform and the limits of fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 594–606
Green B (2021) The flaws of policies requiring human oversight of government algorithms. SSRN Electron J 1–42
Green B, Chen Y (2019) Disparate interactions, pp 90–99. ACM, New York. https://doi.org/10.1145/3287560.3287563

Green B, Chen Y (2019) The principles and limits of algorithm-in-the-loop decision making. In: Proceedings of the ACM on human-computer interaction, vol 3

Green B, Chen Y (2020) Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. arXiv. arXiv:2012.05370

Grgic-Hlaca N, Engel C, Gummadi KP (2019) Human decision making with machine advice: an experiment on bailing and jailing. In: Proceedings of the ACM on human-computer interaction, vol 3

Grgić-Hlača N, Engel C, Gummadi KP (2019) Human decision making with machine assistance: an experiment on bailing and jailing. In: Proceedings of the ACM on human-computer interaction, vol 3, pp 1–25

Hanson RK et al (2017) A five-level risk and needs system: maximizing assessment results in corrections through the development of a common language. https://csgjusticecenter.org/wp-content/uploads/2017/01/A-Five-Level-Risk-and-Needs-System_Report.pdf

Harris GT, Lowenkamp CT, Hilton NZ (2015) Evidence for risk estimate precision: implications for individual risk communication. Behav Sci Law 33:111–127. https://doi.org/10.1002/bsl.2158

Heilbrun K, Dvoskin J, Hart S, Mcniel D (1999) Violence risk communication: implications for research, policy, and practice. Health Risk Soc 1:91–105

Hilton NZ et al (2017) Using graphs to improve violence risk communication. Crim Justice Behav 44:678–694

Hilton NZ, Scurich N, Helmus L-M (2015) Communicating the risk of violent and offending behavior: review and introduction to this special issue. Behav Sci Law 33:1–18. https://doi.org/10.1002/bsl.2160

Howard PD, Dixon L (2012) The construction and validation of the Oasys violence predictor: Advancing violence risk assessment in the English and welsh correctional services. Crim Justice Behav 39:287–307

Karimi-Haghighi M, Castillo C (2021) Efficiency and fairness in recurring data-driven risk assessments of violent recidivism. Proceedings of the ACM Symposium on Applied Computing 994–1002. https://doi.org/10.1145/3412841.3441975

Jahanbakhsh F, Cranshaw J, Counts S, Lasecki WS, Inkpen K (2020) An experimental study of bias in platform worker ratings: the role of performance quality and gender, pp 1–13

Jung S, Pham A, Ennis L (2013) Measuring the disparity of categorical risk among various sex offender risk assessment measures. J Forensic Psychiatry Psychol 24:353–370

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. Q J Econ 133:237–293

Kröner C, Stadtland C, Eidt M, Nedopil N (2007) The validity of the violence risk appraisal guide (vrag) in predicting criminal recidivism. Crim Behav Ment Health 17:89–100

Lee WS, Selart M (2012) The impact of emotions on trust decisions. In: Handbook on psychology of decision-making: new research pp 235–248

Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors J Hum Factors Ergon Soc 46:50–80

Lin ZJ, Jung J, Goel S, Skeem J (2020) The limits of human predictions of recidivism. Sci Adv 6:1–8

Lipkus IM, Samsa G, Rimer BK (2001) General performance on a numeracy scale among highly educated samples. Med Decis Making 21:37–44

Mallari K et al (2020) Do i look like a criminal? Examining how race presentation impacts human judgement of recidivism, pp 1–13. ACM, New York. https://doi.org/10.1145/3313831.3376257. arXiv:2002.01111

McCallum KE, Boccaccini MT, Bryson CN (2017) The influence of risk assessment instrument scores on evaluators' risk opinions and sexual offender containment recommendations. Crim Justice Behav 44:1213–1235

Morgan DL, Krueger RA, King JA (1998) The focus group guidebook. Focus Group Kit. SAGE Publications. https://books.google.es/books?id=5q3k3No59OcC

Mosier KL, Skitka LJ, Heers S, Burdick M (1998) Automation bias: decision making and performance in high-tech cockpits. Int J Aviat Psychol 8:47–63

Portela M, Granell-canut C (2017) A new friend in our Smartphone ? Observing interactions with chatbots in the search of emotional engagement

Rettenberger M, Mönichweger M, Buchelle E, Schilling F, Eher R (2010) Entwicklung eines screeninginstruments zur vorhersage der einschlägigen rückfälligkeit von gewaltstraftätern [the development of a screening scale for the prediction of violent offender recidivism]. Monatsschrift für Kriminologie und Strafrechtsreform 93:346–360

Sambasivan N et al (2021) “Everyone wants to do the model work, not the data work”: data cascades in high-stakes AI. ACM, New York, pp 1–15, USA. https://doi.org/10.1145/3411764.3445518
Scott SG, Bruce RA (1995) Decision-making style: the development and assessment of a new measure. Educ Psychol Meas 55:818–831. https://doi.org/10.1177/0013164495055005017
Scurich N (2015) The differential effect of numeracy and anecdotes on the perceived fallibility of forensic science. Psychiatry Psychol Law 22:616–623
Scurich N, Monahan J, John RS (2012) Innumeracy and unpacking: bridging the nomothetic/idiographic divide in violence risk assessment. Law Hum Behav 36:548–554
Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems.. In: FAT* 2019—Proceedings of the 2019 conference on fairness, accountability, and transparency, pp 59–68
Singh JP, Grann M, Fazel S (2011) A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. Clin Psychol Rev 31:499–513
Skeem J, Monahan J, Lowenkamp C (2016) Gender, risk assessment, and sanctioning: the cost of treating women like men. Law Hum Behav 40:580
Stevenson MT, Doleac JL (2021) Algorithmic risk assessment in the hands of humans
Stevenson M (2018) Assessing risk assessment in action. Minnesota Law Rev 103:303
Storey JE, Watt KA, Hart SD (2015) An examination of violence risk communication in practice using a structured professional judgment framework. Behav Sci Law 33:39–55. https://doi.org/10.1002/bsl.2156
Tan S, Adebayo J, Inkpen K, Kamar E (2018) Investigating human + machine complementarity for recidivism predictions. arXiv. arXiv:1808.09123
van Maanen P-P, Klos T, van Dongen K (2007) Aiding human reliance decision making using computational models of trust, pp. 372–376 (IEEE). https://ieeexplore.ieee.org/document/4427610/
Yin M, Vaughan JW, Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. In: Conference on human factors in computing systems - proceedings, pp 1–12
Yu B et al (2020) Keeping designers in the loop: communicating inherent algorithmic trade-offs across multiple objectives, pp 1245–1257. arXiv:1910.03061
Zhang Y, Liao QV, Bellamy RKE, Vera Liao Q, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: FAT* 2020—Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 295–305. https://doi.org/10.1145/3351095.3372852
Zoe Hilton N, Carter AM, Harris GT, Sharpe AJB (2008) Does using nonnumerical terms to describe risk aid violence risk communication? J Interperson Viol 23:171–188
Zuiderwijk A, Chen YC, Salem F (2021) Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. Gov Inf Q. https://doi.org/10.1016/j.giq.2021.101577

Authors and Affiliations

Manuel Portela[1] · Carlos Castillo[1,2] · Songül Tolan[3] · Marzieh Karimi-Haghighi[1] · Antonio Andres Pueyo[4]

✉ Manuel Portela
manuel.portela@upf.edu

Carlos Castillo
chato@acm.org

Songül Tolan
songueltolan@gmail.com

Marzieh Karimi-Haghighi
marzieh.karimihaghighi@upf.edu

Antonio Andres Pueyo
andrespueyo@ub.edu

[1] Universitat Pompeu Fabra, Campus Poblenou, Barcelona 08018, Spain

[2] Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain

[3] Joint Research Centre, Calle Inca Garcilaso, 3, Seville 41092, Spain

[4] Universitat de Barcelona, Ciutat de Granada, 131, Barcelona 08018, Spain