# People underestimate the errors by algorithms for credit scoring and recidivism but tolerate even fewer errors: A representative study in Germany

PREPRINT – NOT PEER REVIEWED

Felix G. Rebitschek[a,b,1]

Gerd Gigerenzer [a, b]

Gert G. Wagner[a,b,c]

[a]Harding Center for Risk Literacy, Faculty of Health Sciences Brandenburg, University of Potsdam, Potsdam

[b]Max Planck Institute for Human Development, Berlin

[c]German Socio-Economic Panel Study (SOEP), Berlin

[1]Felix G. Rebitschek, **Email:** rebitschek@uni-potsdam.de

**Author Contributions:** F.G.R, G.G. and G.G.W. designed research; F.G.R, G.G. and G.G.W. performed research; F.G.R analyzed data; and F.G.R. and G.G.W. drafted the manuscript; all authors revised the manuscript.

**Competing Interest Statement:** The authors declare no competing interest.

**Classification:** Social Sciences: Psychological and Cognitive Sciences;

**Keywords:** algorithmic decision-making, algorithm acceptance, algorithm errors

**Abstract**

This study provides the first representative analysis of error estimations and error tolerance in a Western country (Germany) with regards to algorithmic decision-making systems (ADM). We examine people's expectations about the accuracy of algorithms that predict credit default, recidivism of an offender, suitability of a job applicant, and health behavior. Also, we ask whether expectations about algorithm errors vary between these domains and how they differ from expectations about errors made by human experts. In a nationwide representative study (N=3,086) we find that most respondents underestimated the actual errors made by algorithms and are willing to tolerate even fewer errors than estimated. Error estimates and error tolerance did not differ consistently for predictions made by algorithms or human experts, but people's living conditions (e.g. unemployment, household income) affect domain-specific tolerance (job suitability, credit defaulting) of misses and false alarms. We conclude that people have unwarranted trust in the competence of ADM systems and evaluate errors in terms of potential personal consequences. Given the general public's low error tolerance, we further conclude that acceptance of ADM appears to be conditional to strict accuracy requirements.

**Significance Statement**

- The general public in Germany overestimates the accuracy of algorithmic decision-making systems (ADM) in predicting recidivism and credit scoring.
- At the same time the public tolerates even fewer ADM errors than actually occur.
- Error tolerance reflects personal conditions and is specific to error types and domains.

## Introduction

This study provides the first representative analysis of error estimations and error tolerance in a Western population (in Germany) with regards to specific algorithmic decision-making (ADM) systems. We examine how accurately algorithms are expected to perform in predicting credit defaulting, recidivism of an offender, suitability of a job applicant, and health behavior.

Algorithmic decision-making (ADM (1)) continues to spread into everyday life. At the same time, claims, risks (2-4), and implementations related to ADM are under debate, e.g. in criminal risk assessment (5-7) or allocation of public resources (8). Despite these controversies, it remains unclear what the general public knows about ADM potential, other than that the concept of algorithms is little understood (9). The public debate centers on laypeople's trust in and fear of algorithms (10). Yet these emotions are not independent of knowledge; they could result from misjudging ADM potential. Overestimating the accuracy of decision support could be associated with unwarranted trust in the competence of algorithms (as opposed to affective trust (11) or trust in motives (12)), while fear could arise from overestimating algorithm failures.

Research on algorithm aversion and appreciation (1, 13) examines both the circumstances under which people trust algorithmic advice (13, 14)—for instance because they perceive the decision problem in question to be objective or to require mechanical skills (15) or because they lack confidence in their own expertise—and the circumstances under which they are mistrustful, for instance (16, 17), in response to slow algorithm responses or to observing algorithm errors. Surprisingly, the expected level of accuracy of algorithms and the perceived competence (12) of algorithmic advice are largely neglected in research on the general population. Fifty-eight percent of Americans expect some level of human bias in ADM systems, and 47% and 49% respectively believe that resume screening of job applicants and scoring for parole are effective (10). Similarly, the Dutch population often evaluates automatic decisions in favor of AI systems (e.g. as being more useful) compared with humans (18). In the present article, we aim to reduce this gap and investigate what the general public expects regarding the accuracy of ADM in the financial, legal, occupational, and health domains. We ask people what they believe are the actual error rates made by ADM and how many errors they consider to be acceptable (1, 17) and check whether their responses meets current ADM accuracy standards (in the financial and the legal domain). To the best of our knowledge, this is the first study comparing error estimates and error tolerance of ADM.

In the US, people's attitudes towards algorithms has been found to be highly context-specific (10). We therefore also examine whether the degree of error tolerance is associated with personal conditions, e.g. a risk-seeking preference (19). Classifying ADM systems can balance two different types of errors, misses and false alarms, each associated with different consequences (costs). Thinking about types of decision errors and related costs can affect tolerance of errors (e.g. 'bias' in signal detection theory (20); cost-sensitive error management (21)), and this tolerance may differ in the legal (22) and in the medical domain (23). From the perspective of an unemployed person, for instance, mistakenly being overlooked for a job by an ADM (a miss) is likely more costly than being hired in spite of being unsuitable (a false positive). In our analysis, we therefore relate assessed error tolerance to critical factors such as unemployment phases. Also, we compare error tolerance with the typical error preference of exemplary stakeholders (e.g. non-recidivating offenders want to avoid a false alarm). We additionally explore factors that may influence algorithm error tolerance and underlying attitudes towards technology, such as risk preference (24), gender (9, 25), and age (26). For instance, only one third of US-Americans above 50 years of age compared with half of those 18 to 29 years of age believe that algorithms can be free of human biases (10).

Finally, we compare error estimation and tolerance regarding algorithmic and expert advice across domains. Earlier studies indicate that in the medical domain, people prefer human experts over automated care because, among others, they believe that individual circumstances are neglected by machines (27). People were also found to trust computers less than physicians to make good recommendations (28) and, crucial for error perception, to consider algorithm-based advice as being less accurate than a clinicians' advice (29). In the domain of people analytics, human interviewers are also assumed to be more accurate and useful than algorithmic decision-aids (30). At least for those two domains, the few existing studies thus indicate that people perceive algorithms to make more errors than human experts do. On the other hand, bestseller authors and commercial companies have aggressively promoted "AI"—which in some instances has demonstrated better performance for selective tasks (e.g. image classification(31) and deceptive text detection(32))—as being superior to human experts by underscoring the accomplishments of algorithms based on big data that have, for instance, beaten the best Go players and, as in the case of IBM's supercomputer Watson, promise to revolutionize health care. From that point of view, "it would be madness not to follow their advice" (33).

We conducted a survey of a large population-representative sample of members of private households in Germany (by means of the Innovation Sample of the German Socio Economic Panel

Study, SOEP IS). In the survey, 3,086 respondents were questioned about estimated and accepted error rates in current credit scoring (introductory task) before being randomly assigned to decision scenarios either on algorithms or on experts (between-subjects). Respondents were requested to provide estimated and accepted error rates in a people analytics problem (predicting a suitable job candidate), a legal problem (predicting recidivism of an offender), and a health problem (evaluating health behavior). For instance, respondents were asked with respect to ADM and recidivism: "In the US judiciary, offenders are regularly reviewed for early release. The computer program COMPAS assesses whether a criminal will recidivate and commit another crime within the next 2 years. Now please imagine a group of 100 offenders who are actually at risk of recidivism: How many of them do you think the computer program incorrectly assesses as not being at risk of recidivism [estimated miss rate]? What would you personally find acceptable: How many of these 100 offenders at most could be wrongly assessed as not being at risk of recidivism by the computer program [accepted miss rate]? Now please imagine a group of 100 offenders who are not at risk of recidivism: How many of them do you think the computer program wrongly considers to be at risk of recidivism [estimated false alarm rate]? What would you personally find acceptable: How many of these 100 offenders at most could be wrongly estimated to be at risk of recidivism by the computer program [accepted false alarm rate]?"

## Results

### *Estimated and accepted error rates of ADM systems*

Across domains and types of errors, about one in every four algorithmic decisions was estimated to be wrong (Figures 1A-D), with the lowest estimates of errors for credit scoring (Figure 1D). Only a few respondents believed in perfect algorithm (or expert) performance (Table S1).

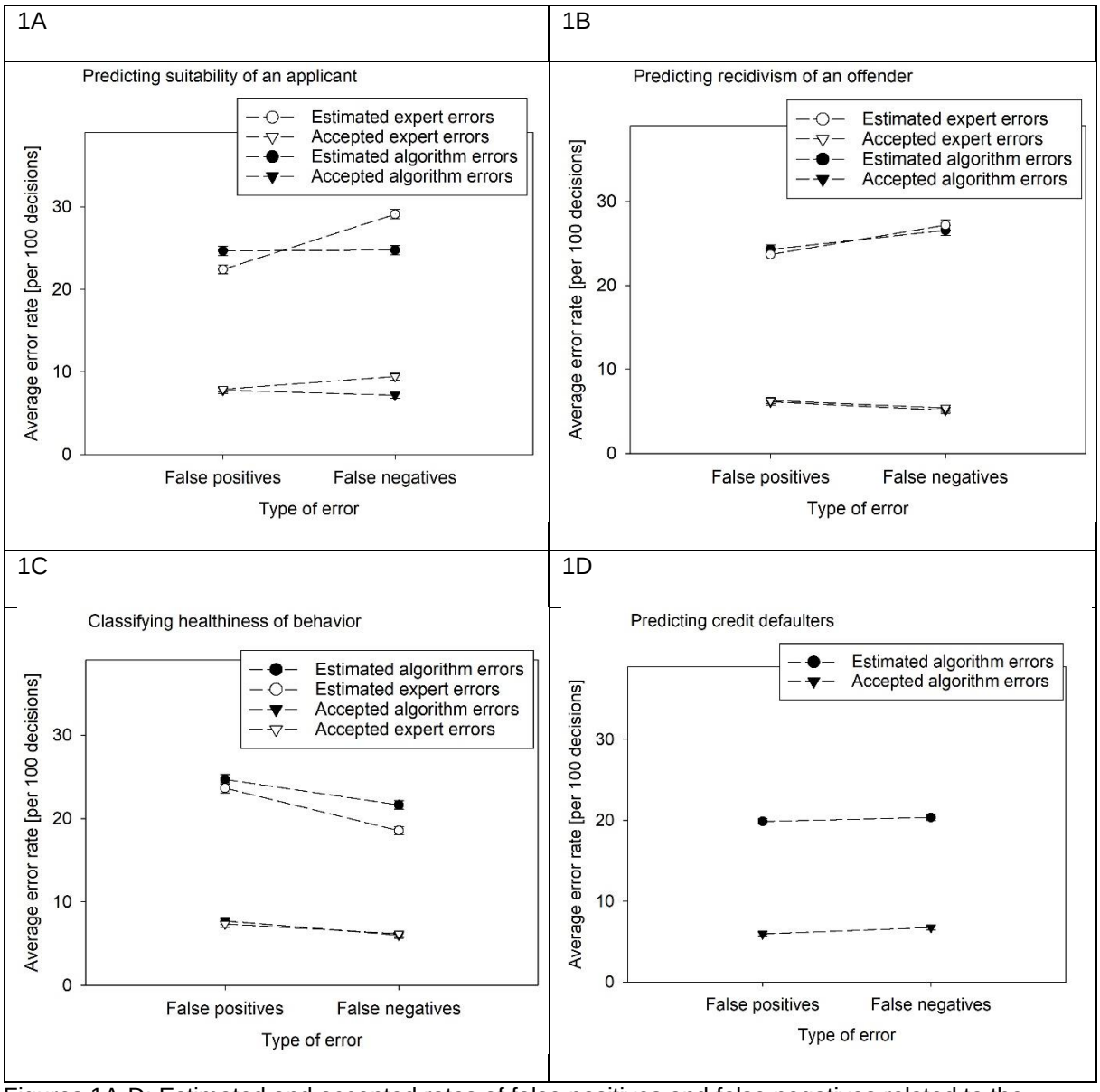
Minimally more false negatives than false positives were estimated for the algorithmic prediction of recidivism ($F(1,1314) = 15.75$, $P < .001$, $\eta_p^2 = 0.01$), but false negatives were considered less acceptable ($F(1,1321) = 4.32$, $P = .038$, $\eta_p^2 < 0.01$). For the algorithmic evaluation of health behavior (Figure 1C), more false positives (falsely healthy) than false negatives were estimated ($F(1,1297) = 39.17$, $P < .001$, $\eta_p^2 = 0.03$) and considered acceptable ($F(1,1310) = 22.76$, $P < .001$, $\eta_p^2 = 0.02$).

Notably, there is a consistent and large difference between the estimated and maximally accepted rates of algorithm error across all four domains. Expressed in percentage points, the differences in false positives (FP) and false negatives (FN) are for application suitability ($\Delta_{FP} = 17$, $\Delta_{FN}$

142 = 17), for recidivism ($\Delta_{FP}$= 18, $\Delta_{FN}$ = 21), for health behavior ($\Delta_{FP}$ = 17, $\Delta_{FN}$ = 16), and for credit scoring

143 ($\Delta_{FP}$ = 14, $\Delta_{FN}$ = 14).

144




145 Figures 1A-D: Estimated and accepted rates of false positives and false negatives related to the

146 assessments of job suitability (1A), recidivism (1B), and health behavior (1C) by algorithms and

147 experts, and to credit scoring (1D). Data are weighted for representativeness. Error bars show the

148 standard errors of the mean. For predicting credit defaulters, estimates of expert errors were not

149 elicited.

150

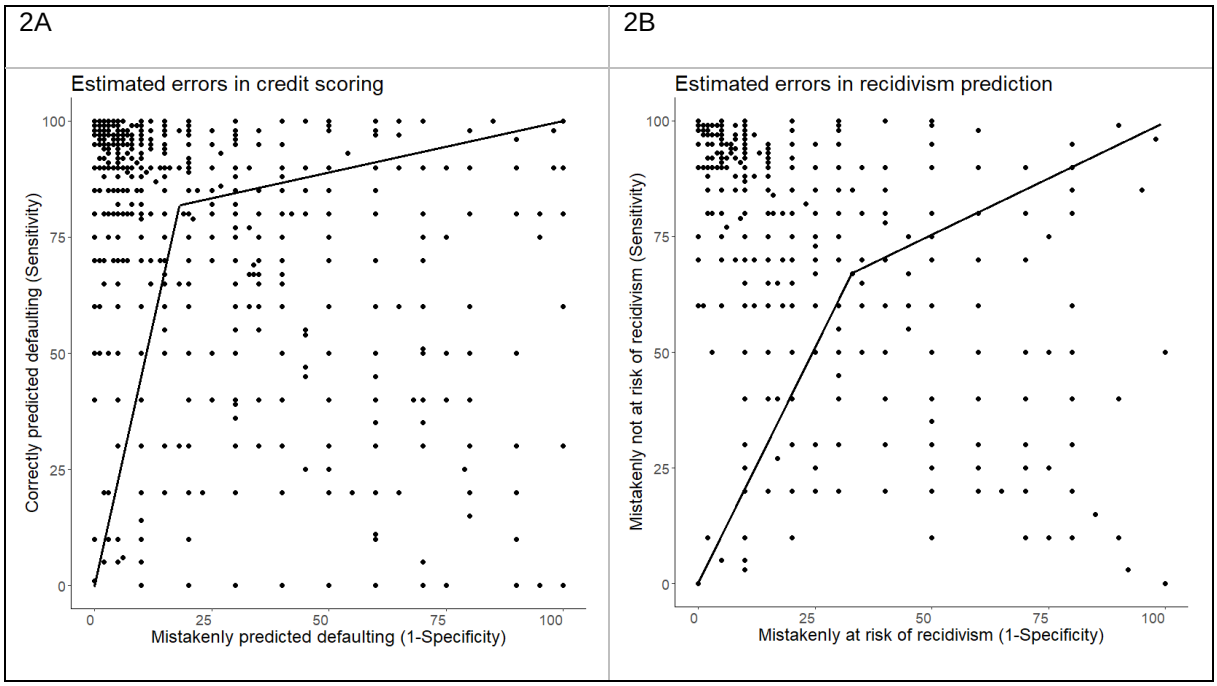
151

152

***Estimated and actual errors of ADM systems***

To assess whether estimated algorithm errors correspond to actual errors, we compared respondents' estimates with actual performance metrics for recidivism and credit scoring algorithms (for the scenarios of health behavior and people analytics, reliable metrics are not publicly available). For credit scoring, most respondents' estimated error rates (false positives and false negatives) were substantially lower than those reported by the German SCHUFA, which is the major credit scorer in Germany. SCHUFA reports an accuracy of .81 of its (banking) credit score in terms of the value of the area under the receiver operating curve (AUC) (34). This is lower than the unweighted AUC of 0.90 that we calculated based on median sensitivity (.90) and specificity (.90) of respondents' estimates. Figure 2A (left) shows each of the respondents' estimates for the false negative rate and false positive rate, expressed as sensitivity (1 – false negative rate) and 1 - specificity (false positive rate). Each dot corresponds to one respondent. The vast majority of dots are in the top left corner, that is, show a very high estimated sensitivity (low miss rate) and a low false alarm rate. The kinked line corresponds to the actual AUC as reported by the SCHUFA. Given that the SCHUFA does not provide data for the entire curve, we represent the value of .81 as a best case with assumed error balance. This is shown with the kinked line that is symmetric around the top-left to bottom-right diagonal. Relative to this line, the points above represent overestimation and the points below represent underestimation of the accuracy of the credit scoring algorithm.

The same analysis is shown for the prediction of recidivism in Figure 2B (right). The accuracy of the widely used COMPAS algorithm (General Recidivism Risk Scale) is lower than that of the credit scoring algorithms, with an AUC of .68 (5). Again, the vast majority of the respondents overestimated its accuracy. Respondents' median estimated sensitivity (.80) and specificity (.80) correspond to an unweighted AUC of .80. In other words, the performance of the algorithm does not meet the population's expectations as expressed in estimations and acceptance. We characterize this gap, and to whom it is relevant, in the following section.



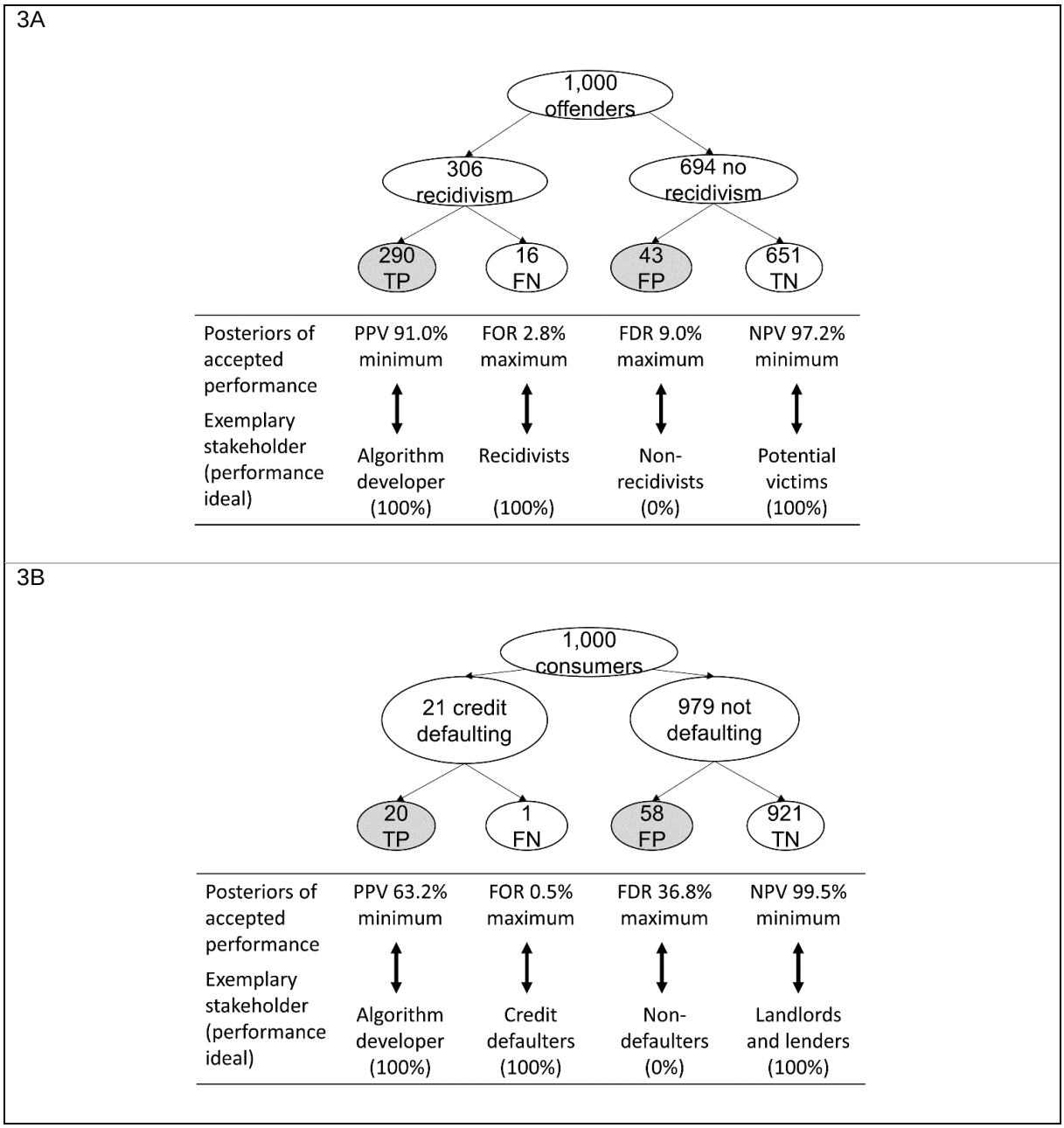

183 Figure 2: Respondents vastly overestimate the accuracy of algorithms for credit scoring and recidivism
184 prediction. The scatterplots show each respondent's estimates of the false positive rate (1 - specificity)
185 and false negative rate (shown as its complement, the sensitivity) in credit scoring (*n* = 2,740) and in
186 recidivism prediction (*n* = 1,325). Each point corresponds to one respondent. The kinked line shows
187 simplified ROC curves for current credit scoring accuracy (SCHUFA Credit Score algorithm,
188 AUC=0.81) and for recidivism prediction (COMPAS algorithm, AUC=0.68).

189

190 ***Acceptance of different errors made by ADM systems***

191 To interpret tolerated errors, their base rates have to be taken into consideration. What counts for an
192 individual decision are the posterior probabilities (how likely a certain prediction is correct or incorrect),
193 not the miss rates and false alarm rates. In a first step, by linking the rates of accepted algorithms
194 errors (weighted for representativeness) and the reported base rates of 30.6% for recidivism (within
195 four years, any offense, both genders; Table 2 in (35)) and of 2.1% for credit defaulting (within one
196 year (36)), we calculated the implied posterior probabilities. As the Figures 3A-B) show, the accepted
197 minimum positive predictive value (PPV) for recidivism is about 91%, whereas that of credit defaulting
198 is only 63%, despite the credit algorithm being more accurate than the one for recidivism.

199




Figures 3A-B: Natural frequency trees (37) of algorithm error acceptance for base rates of 30.6% recidivists (3A) and of 2.1% credit defaulters (3B), and comparison of minimum performance acceptance (posterior probabilities) of respondents with performance ideals of exemplary stakeholders. Posteriors shown are the acceptable minimum PPV (positive predictive value; e.g. that a person will commit another crime if testing positive), minimum NPV (negative predictive value: e.g., that a person will not commit another crime if testing negative), maximum FOR (false omission rate = 1-NPV), and maximum FDR (1 - PPV); TP (true positive), TN (true negative), FP (false positive), FN (false negative). Note that depicted averages of individual posteriors are not equal to theoretical posteriors based on the depicted aggregated TPs, FPs, TNs, and FNs.

In a second step, we compared the calculated posterior probabilities of error acceptance with ideal posteriors according to different stakeholders' goals (Figures 3A-B). For algorithm developers, the key goal is increasing PPV (precision) in the direction of 100% (38). An offender at risk of recidivism, in contrast, wants to falsely test negative and be released, that is, a test with a high FOR. In contrast to offenders at risk, those not at risk have an interest in not being falsely diagnosed as at risk of recidivism (low FDR). Finally, potential victims of defenders on bail will demand tests that maximize NPV towards 100%, that is, the probability that offenders will not commit another crime if they are tested negative and are released. The comparison showed that people's preferences for recidivism prediction overlap with performance ideals of developers, of potential victims, and of non-recidivating offenders, but not with recidivating offenders: False negatives were considered less acceptable than false positives.

Figure 3B provides similar examples of conflicting interests for predicting default. For instance, landlords and lenders do not want to contract with defaulting consumers; thus, their goal is to reach an NPV that approaches 100%. Customers at risk of default, in contrast, profit from tests that have low NPV, which is equivalent to high FOR. People's error tolerance did partially match performance ideals of the developers. Of note, our sample signaled substantial dissimilarity with non-defaulters because many false positives, and more of them than false negatives, were considered acceptable.

Based on the insight that posteriors in the interest of multiple stakeholders can only be partially achieved, it is examined in the following how error tolerance in general could depend on personal consequences from errors in specific domains.

***Factors of domain-specific error acceptance***

A linear regression ($F_{FP}(6, 1124) = 52.79$, $P < .001$, $R_{corr}^2 = 0.22$) that controlled for error estimations and age-, income-, and gender-specific differences in accepting algorithms' and experts' errors showed that reporting phases of unemployment in the past ten years (std. $\beta = 0.10$) went along with generally increased acceptance of falsely assessed suitability. In contrast, overlooking suitable job candidates ($F_{FN}(6, 1125) = 49.35$, $P < .001$, $R_{corr}^2 = 0.20$) was less accepted by men (std. $\beta = -0.09$) and by older people (std. $\beta = -0.08$).

Furthermore, linear regressions ($F_{FN}(5, 1685) = 29.93$, $P < .001$, $R_{corr}^2 = 0.08$) revealed that misses (false negatives) of recidivating offenders were less accepted by younger respondents (std. $\beta_{FN} = 0.08$), those with higher income (std. $\beta_{FN} = -0.07$), and women (std. $\beta_{FN} = -0.08$). Similar

relationships were not observed for false positives, with $F_{FP}(5, 1669) = 40.66$, $P < .001$, $R_{corr}^2 = 0.11$, and generally not in the case of the evaluation of health behavior ($F_{FP}(8, 1625) = 42.39$, $P < .001$, $R_{corr}^2 = 0.17$; $F_{FN}(8, 1635) = 66.43$, $P < .001$, $R_{corr}^2 = 0.24$), even if controlling for body mass index, health, and insurance status.

With regards to errors in credit scoring, fewer false negatives were accepted ($F(5, 1733) = 115.03$, $P < .001$, $R_{corr}^2 = .25$) by older respondents (std. $\beta_{FP} = -0.05$) and those reporting higher household income (std. $\beta_{FP} = -0.07$). The latter also accepted fewer false positives (std. $\beta_{FP} = -0.05$), ($F(5,1738) = 88.10$, $P < .001$, $R_{corr}^2 = .20$).

In sum, laypeople take into account potential benefits and harms of certain types of errors, regardless of who or what caused the error. In the last step, we examined the differences in estimations and acceptance of errors made by ADM systems and experts.

***Experts vs. algorithms***

We analyzed whether estimated and accepted errors vary for ADM and human experts (Table S2). A 3x2x2 ANOVA (advisor x type of error x domain) indicated no general difference ($F(1,2446) = 0.14$, $P = .707$) between expert and algorithm error estimations but that these differed depending on the domain ($F(1,4892) = 11.79$, $P < .001$, $\eta_p^2 = 0.01$) and on the type of error within the domain ($F(2,4892) = 32.81$, $P < .001$, $\eta_p^2 = 0.01$). For algorithms, (4.6 percentage points) lower rates of false negatives were estimated in the case of suitable applicants ($F(1,2622) = 78.50$, $P < .001$, $\eta_p^2 = 0.03$). However, minimally higher error rates were estimated for recognizing healthy behavior, where false positives and false negatives are estimated to be 1.1 and 2.0 percentage points higher, respectively, for ADM than for humans ($F(1,2567) = 8.87$, $P < .001$, $\eta_p^2 < 0.01$).

Differences in acceptance of errors made by algorithms or experts depended on the domain ($F(1,4896) = 3.08$, $P = .046$, $\eta_p^2 < 0.01$) and on the type of error within the domain ($F(1,4896) = 3.45$, $P = .032$, $\eta_p^2 = 0.01$) (Table S2). For suitability prediction, there was a generally lower acceptance of algorithm errors in suitability prediction ($F(1,2631) = 4.92$, $P = .027$, $\eta_p^2 < 0.01$); false negatives were also considered less acceptable than false positives in the case of algorithms ($F(1,2631) = 12.06$, $P = .001$, $\eta_p^2 = 0.01$). For recidivism prediction, in contrast, differences in acceptance are limited. False positives seem to be considered less acceptable when caused by algorithms than by experts. Mistakenly predicting false recidivism was generally less accepted ($F(1,2780) = 4.24$, $P = .040$, $\eta_p^2 <$

0.01). Notably, it was generally considered less acceptable to falsely diagnose unhealthy behavior than to overlook it ((*F*(1,2575) = 37.63, *P* < .001, $\eta_p^2$ = 0.01)).

What factors contribute to differences in evaluating the performance of an algorithm and an expert? Generally, age (std. $\beta$ = -0.10, *P* < .001) and household income (std. $\beta$ = -0.08, *P* = .007) but not gender and risk preference (Figure S1), if controlled for estimated error rates (std. $\beta$ = 0.55, *P* < .001), were negatively related with acceptance of ADM errors (*F*(5,933) = 89.03, *P* < .001, $R_{corr}^2$ = 0.32). However, the pattern was similar for errors made by human experts.

An analysis of factors that lead to ADM-specific acceptance of errors confirms that people above the age of 29 (((*F*(3,2717) = 9.01, *P* < .001, $\eta_p^2$ = 0.01)) accept a higher number of human than algorithm errors (about 1 to 2 more per 100 predictions), while those younger than 29 accept fewer human errors and more algorithm errors (a difference of about 4 in 100 predictions). Similar distinctions could not be confirmed for gender ((*F*(1,2721) = 0.20, *P* = .659) or for monthly household net income (*F*(4,1752) = 1.82, *P* = .122).

Finally, an additional analysis of the relationship between estimations and acceptance, which were strictly correlated (between r = .27 and .43), did not indicate differential developments of error acceptance with increasing perception of errors by algorithms and experts (Figure S2).

**Discussion**

Our nationwide representative study in Germany delivered two major results. First, the vast majority of respondents underestimated error rates for credit scoring and recidivism prediction. Second, respondents tolerated even fewer algorithm errors than they expected to occur, only 5% to 7% errors compared with their estimates of about 20% to 30% errors. This gap between tolerated error rate and both the perceived and actual accuracy of algorithms is striking. These findings indicate a lack of public trust in the *competence* of algorithms (confidence), which depends on knowledge about past performance. This is in line with the concerns of the American public about the acceptability of algorithms in recidivism prediction and resume screening of job applicants: More than 50% find them unacceptable (10). For one, they perceive those systems to be incapable of doing nuanced work across individuals.

The lack of trust in the competence of algorithms needs to be distinguished from a lack of trust in the *intentions* of those behind the algorithms (12), which is the common target of digitalization campaigns, such as in the EU White Paper on Artificial Intelligence (39). Our result suggests that

winning trust in ADM requires better accuracy of ADM systems, even beyond the performance overestimations of our respondents. Because the respondents accepted fewer errors by ADM than by human experts in some domains, the public may expect some ADM systems to be more error-free than humans in order to tolerate them. ADM was estimated to perform roughly at par with experts. The only exceptions were estimations that on the one hand algorithms detect suitable job candidates better than experts and that on the other hand experts (physicians) are perceived to make slightly fewer errors of both types in evaluating health behavior.

Anticipated consequences (e.g. costs) of different types of error appear to inform judgments about error tolerance: People who experienced phases of unemployment in the past ten years are more open for errors in their favor, namely false positives in predicting suitable job candidates. People with higher household income tolerate fewer errors in credit scoring. A higher tolerance of false positives than false negatives in predicting defaulting is noteworthy, given that 98% of the people in Germany are non-defaulters in the credit scoring system discussed. Given that the proportion of lenders, landlords, and entrepreneurs among the population is quite limited, future research should ask: Do people who tolerate false positives fear harms to the economy if algorithms were not that sensitive or do they simply ignore the personal consequences of being classified falsely as a defaulter?

*Research implications*

The general public's perception of the accuracy of algorithms and its general tolerance of ADM error have been a neglected issue in previous research, which has focused on issues such as people's opinions on discrimination, safety, and surveillance by algorithms, often implying the existence of high accuracy. Our representative analysis of error estimations and error tolerance of the German population shows that the accuracy of real ADM systems in credit scoring and recidivism prediction is too low from the public's perspective. Observed as well as labelled accuracy of predictive algorithms (40, 41), even if it is low (42), and also confidence information for each prediction (43) can improve trust in algorithms (although are users trusting independent of performance (44)). Contrary to studies that identify risk-seeking preferences to explain the preference of imperfect algorithms over human advice (19), our representative study across three domains did not find any indications of a relationship between risk preference and the acceptance of algorithm errors.

Similar to acceptance, the accuracy of predictive algorithms can often be improved by making them simpler (29). Simplicity implies transparency and understandability (e.g. explaining a prediction (42)), which are potential measures to increase trust in ADM systems (45). The design of such systems may take successful human strategies of decision making under uncertainty in the real world as a starting point (46). This is of particular importance because merely using interpretable models (47) and explaining complex models (48) does not necessarily form trust in ADM systems. A further avenue lies in equipping regulators with tools that enable lay-comprehensible algorithm assessments in terms of consequences, performance, and fairness. This calls for research that combines trust in competence, such as in the accuracy of algorithms, with trust in motivation, such as transparency.

*Educational implications*

Currently, ADM systems are spreading through most life domains, including personalized medicine, consumer scoring, credit scoring, insurers, people analytics, learning analytics, and predictive policing. This cultural change requires citizens to acquire digital risk literacy (49), including the ability to evaluate the performance of predictive algorithms. Accordingly, schools need to teach students the requisite concepts to understand ADM systems, including input, output, consequences, features, feature weights, fairness, and performance-related properties such as false alarm rate, miss rate, positive predictive value, cross-validation, along with the skills to locate the relevant information.

Yet people should not only learn about the actual performance that is delivered by credit scoring and recidivism prediction. With regards to both algorithms and experts it is important to keep in mind that people are sensitive to types of error because these present different costs for them and others. Laypeople consider these potential costs when confronted with algorithmic decision support. Educational materials on decision-support systems should thus inform readers about the performance of algorithms. Facilitating this understanding, however, also requires a culture of transparency and trust created by commercial firms or imposed by regulators, where people gain easy access to relevant information about the reliability of algorithms in credit scoring, people scoring, and other domains (50).

*Limitations*

One limitation to the present study is that the estimates for the actual accuracy of the credit scoring algorithm are self-reported by the SCHUFA. The accuracy estimates for COMPAS, in contrast, stem

from independent scientific studies. The SCHUFA is a commercial company, and one could argue that their estimate of their own algorithms might be inflated. We have no evidence of this possibility, but if it were the case, that would mean that the public overestimates the accuracy of the algorithm to an even greater extent than what we analyzed in Figure 2. In other words, the gap between estimated accuracy and actual accuracy would be even larger.

Nor can we exclude the possibility that people in Germany have a better sense of the accuracy of other ADM systems, as we could only compare respondents' estimates to the available accuracies for the domains of credit scoring and recidivism prediction. However, as long as there are no further learning possibilities or transparent information available for other ADM systems, we have no reason to assume that the overestimates are unique to recidivism and credit scoring.

## Materials and Methods

Data. The study data were collected as a part of the "Innovation Sample" of the German Socio-Economic Panel (SOEP IS) (51). For this longitudinal survey, which started in 2011, all adult and youth members of the same households are interviewed every year. The panel's representativeness relies on interviewing split-offs (younger panel members founding their own families). Our household sample with 3,086 respondents (52.1% female, M=50.4 years of age (SD = 19.0)) is representative of the adult population in Germany (52).

Design. In accordance with an experimental between-subjects design, after responding to credit scoring items, respondents were randomly assigned to either algorithm or expert predictions (recidivism, job suitability, healthiness of behavior).

Measures. We used four partially known scenarios to introduce ADM systems, of which credit scoring is fully established in Germany, recidivism prediction is established outside Germany, and suitability prediction is currently being established in Germany; the evaluation of healthy behavior was derived from established health bonus programs.

Error rates for each algorithm and each expert were assessed with a normalized frequency format. The following is an item example for overlooking recidivism in algorithms: "In the US judiciary, offenders are regularly reviewed for early release. The computer program COMPAS assesses whether a criminal will recidivate and commit another crime within the next 2 years. Now please imagine a group of 100 offenders who are actually at risk of recidivism: How many of them do you think the computer program incorrectly assesses as not being at risk of recidivism?"; "What would you

personally find acceptable: At most, how many of these 100 criminals could be wrongly assessed as not being at risk of recidivism due to the computer program?“ A validated self-report item about risk attitude (24) was included (11-point). Standard demographic variables in the panel that we used were gender, age, household income (per month), event of unemployment (in the last ten years), health insurance status (statutory or private), and self-rated health status (5-point).

Data were analyzed with regression techniques. All data were weighted on the person level to fine-tune for representativeness. All data (at the German Socio Economic Panel) and materials (at PNAS) are freely available.

Acknowledgments

Support for this work was provided by the German Socio Economic Panel Study (https://www.diw.de/en/diw_01.c.390440.en/soep_is.html) and by a Max Planck Society Fellowship (to G.G.W.). Furthermore, F.G.R. was supported through the RisikoAtlas project (www.risikoatlas.de), which was funded by the Federal Ministry of Justice and Consumer Protection (BMJV) on the basis of a resolution of the German Bundestag via the Federal Office for Agriculture and Food (BLE) within the framework of the Innovation Programme. We are grateful to Björn Meder for his advice on data presentation. We thank David Richter (SOEP) and his team as well as the team of the fieldwork organization Kantar Munich for implementing our survey questions in SOEP IS 2018/19. We thank our proofreader Rona Unrau for her critical eye.

The study was approved by the Institutional Ethics Board of the Max Planck Institute for Human Development, Berlin (Germany). It was carried out in accordance with the guidelines and regulations of the Max Planck Society for the Advancement of Science in Germany.

## References


1. Burton JW, Stein M-K, & Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* n/a(n/a).
2. Russell SJ (2019) *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin).
3. Smith BC (2019) *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press).
4. Angwin J, Larson J, Mattu S, & Kirchner L (2016) Machine bias. *ProPublica, May* 23:2016.
5. Dressel J & Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1):eaao5580.
6. Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, & Mullainathan S (2018) Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1):237-293.
7. Stevenson MT & Doleac JL (2019) Algorithmic Risk Assessment in the Hands of Humans. *Available at SSRN*.
8. Lohninger T & Erd J (2019) SUBMISSION for the report to the UN General Assembly on digital technology, social protection and human rights. (Vienna).
9. Grzymek V & Puntschuh M (2019) What Europe Knows and Thinks About Algorithms Results of a Representative Survey. Bertelsmann Stiftung eupinions February 2019.
10. Smith A (2018) Public attitudes toward computer algorithms. in *Pew Research Center* (Pew Research Center).
11. McAllister DJ (1995) Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal* 38(1):24-59.
12. Twyman M, Harvey N, & Harries C (2008) Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making* 3(1):111.
13. Logg JM, Minson JA, & Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90-103.
14. Castelo N, Bos MW, & Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809-825.
15. Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5(1):2053951718756684.
16. Efendic E, van de Calseyde P, & Evans A (2019) Slow decision speed undermines trust in algorithmic (but not human) predictions. *PrePrint*.
17. Dietvorst BJ, Simmons JP, & Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114-126.
18. Araujo T, Helberger N, Kruikemeier S, & De Vreese CH (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society* 1-13.
19. Jay Dietvorst B & Bharti S (2019) Risk Seeking Preferences Lead Consumers to Reject Algorithms in Uncertain Domains. in *ACR North American Advances*, eds Bagchi R, Block L, Lee L, & Duluth (Association for Consumer Research), pp 78-81.
20. Green DM & Swets JA (1966) *Signal detection theory and psychophysics* (Wiley, New York).
21. Haselton MG & Buss DM (2000) Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology* 78(1):1-11.
22. Mitchell G & Garrett BL (2019) The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence. *Behavioral Sciences & the Law* 37(2):195-210.
23. Shiloh S (2010) An Experimental Investigation of the Effects of Acknowledging False Negative and False Positive Errors on Clients' Cancer Screening Intentions: The Lesser of Two Evils? *Applied Psychology: Health and Well-Being* 2(2):204-221.

24. Frey R, Pedroni A, Mata R, Rieskamp J, & Hertwig R (2017) Risk preference shares the psychometric structure of major psychological traits. *Science advances* 3(10):e1701381.
25. Pierson E (2017) Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.
26. Mossberger K, Tolbert CJ, & Stansbury M (2003) *Virtual inequality: Beyond the digital divide* (Georgetown University Press).
27. Longoni C, Bonezzi A, & Morewedge CK (2019) Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46(4):629-650.
28. Promberger M & Baron J (2006) Do patients trust computers? *Journal of Behavioral Decision Making* 19(5):455-468.
29. Eastwood J, Snook B, & Luther K (2012) What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making* 25(5):458-468.
30. Diab DL, Pui S-Y, Yankelevich M, & Highhouse S (2011) Lay Perceptions of Selection Decision Aids in US and Non-US Samples. *International Journal of Selection and Assessment* 19(2):209-216.
31. He K, Zhang X, Ren S, & Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, pp 1026-1034.
32. Ott M, Choi Y, Cardie C, & Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, (Association for Computational Linguistics), pp 309-319.
33. Harari YN (2016) *Homo Deus: A brief history of tomorrow* (Random House).
34. SCHUFA (2019) Zuverlässiger Score. Sichere Bank. - Der Schufa Score für Banken 3.0.
35. Brennan T, Dieterich W, & Ehret B (2009) Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36(1):21-40.
36. SCHUFA (2019) Kredit Kompass 2019.
37. Gigerenzer G & Hoffrage U (1995) How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review* 102(4):684-704.
38. Demartini G & Mizzaro S (2006) A classification of IR effectiveness metrics. *European Conference on Information Retrieval*, (Springer), pp 488-491.
39. Ciucci M & Gouardères F (2020) The White Paper on Artificial Intelligence.
40. Yin M, Wortman Vaughan J, & Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp 1-12.
41. Yu K, Berkovsky S, Taib R, Zhou J, & Chen F (2019) Do I trust my machine teammate? an investigation from perception to decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp 460-468.
42. Lai V & Tan C (2019) On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp 29-38.
43. Zhang Y, Liao QV, & Bellamy RK (2020) Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *arXiv preprint arXiv:2001.02114*.
44. Springer A, Hollis V, & Whittaker S (2017) Dice in the black box: User experiences with an inscrutable algorithm. *2017 AAAI Spring Symposium Series*.
45. Ribeiro MT, Singh S, & Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 1135-1144.
46. Hafenbrädl S, Waeger D, Marewski JN, & Gigerenzer G (2016) Applied decision making with fast-and-frugal heuristics. *Journal of Applied Research in Memory and Cognition* 5(2):215-231.

47. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, & Wallach H (2018) Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.
48. Cheng H-F*, et al.* (2019) Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp 1-12.
49. Gigerenzer G (2015) *Risk savvy: How to make good decisions* (Penguin).
50. O'Neill O (2018) Linking Trust to Trustworthiness. *International Journal of Philosophical Studies* 26(2):293-300.
51. Richter D & Schupp J (2015) The SOEP Innovation Sample (SOEP IS). *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift für Wirtschafts-und Sozialwissenschaften* 135(3):389-400.
52. Rohrer JM, Egloff B, & Schmukle SC (2015) Examining the effects of birth order on personality. *Proceedings of the National Academy of Sciences* 112(46):14224-14229.