Algorithm Audit

Advice document

Predicting irresponsible driving

Algoprudence
ALGO:AA:2025:01:A

July 2025

Key takeaways normative advice commission

> **Model validity is fundamental**
The algorithm must be altered to specifically predict driving behavior that causes damage, not general platform misuse. As for any risk prediction model, getting alignment between training data and intended purpose is a critical prerequisite.

> **Balance monitoring with user autonomy**
Monitoring irresponsible driving to reduce damage costs is a legitimate business interest but must not become excessive surveillance or veer into paternalistic advice about general driving habits.

> **Meaningful transparency required**
Users need specific explanations about what driving behavior triggered the warning and clear guidance for improvement, not generic warnings or confusing technical jargon that means nothing to the average driver.

> **Careful variable selection**
Speeding has obvious safety implications, but acceleration and similar variables are trickier. They depend on context and may just reflect personal driving preferences. Before including them, there must be solid evidence linking them to actual damage risk, not just different driving styles or environments.

> **Human oversight essential**
Human analysts currently override 50-60% of the model's recommendations, demonstrating real discretion rather than rubber-stamping. This meaningful human oversight must continue.

Summary advice

The commission judges that algorithmic risk prediction for identifying irresponsible driving behavior should take place under strict conditions and should be weighed against alternative methods of reducing damage. The validity of the prediction model is a critical prerequisite, and hence the current mismatch between the stated objective (predicting irresponsible driving) and the target variable in training (user bans for a wide variety of misuse) must first be resolved. The commission emphasizes that while monitoring to reduce damage cost may be a legitimate business interest, it should not become excessive surveillance or be used for paternalistic feedback on users' general driving style. Users should receive specific, meaningful explanations about which driving behaviors triggered warnings, not generic notifications or lists of technical variables that users cannot comprehend. Variable selection must be carefully justified, with speeding as the most legitimate variable, while contextual behaviors like fast acceleration or hard braking require attention to driving context and solid evidence in what sense they are related to damage risk. The commission recommends maintaining substantial human review of algorithmic recommendations, to mitigate the risk that warnings are unduly sent and to facilitate appeal and redress by users.

Table of Contents

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are four-fold:

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see our white papers and public standards.

Normative advice commissions

Advising on ethical issues that arise in concrete algorithmic practice through deliberative and diverse normative advice commissions, resulting in algoprudence

Technical tools

Implementing and testing technical tools to detect and mitigate bias, e.g., sociotechnical evaluation of generative AI, unsupervised bias detection and synthetic data generation.

Project work

Supporting public and private sector organisations with specific questions regarding responsible use of AI, from a not-for-profit perspective.

1. Preface

This advice is the result of the deliberation by an independent normative advice commission. Algorithm Audit has drafted this advice based on a discussion had during a physical meeting of the advice commission. During this meeting, several ethical questions regarding the use of algorithmic risk profiling to predict irresponsible driving behavior were discussed.

The specific case on which this advice is based on a risk profiling model used for this purpose by a business-to-consumer car sharing platform. Through machine learning a risk model (balanced random forest) is trained to identify driving behavior associated with risk and vehicle damage. For each user the model calculates a risk score following each new trip. If a risk score surpasses a threshold, a warning is sent to the user. If their driving behavior does not improve in subsequent trips, the platform may block the user from its services following a human review.

This case was chosen for review by Algorithm Audit because it serves as a clear example of how machine learning-based risk profiling is applied in real-world contexts, such as e-commerce, banking, and human resources – fields where details about these methods are rarely disclosed by organizations. For this case, the car sharing platform provided detailed information about training data, hyperparameters used in its balanced random forest algorithm and the interplay between algorithmic-driven decisions and human oversight when blocking users. All specifics are available in the problem statement 'Predicting irresponsible driving behavior' (ALGO:AA:2025:01:P).

By this case Algorithm Audit expands upon its previous investigations of algorithmic risk profiling in the private sector. While an earlier case study (ALGO:AA:2022:01:A) investigated an e-commerce platform's risk model which raised concerns about proxy discrimination through apparently innocuous variables like SIM card type, this case is compelling because it presents fundamentally different ethical considerations. The e-commerce case centered primarily on protecting business interests against payment fraud. In contrast, risk profiling of irresponsible driving affects not only business operation costs but also public road safety. Moreover, in this case the ethical considerations go beyond concerns about (proxy) discrimination vis-a-vis protected legal grounds. A primary concern is the monitoring of driving behavior itself and what variables can justifiably be used for predicting damage risk. The combination of rich technical information and unique context of application makes this case especially valuable for independent and deliberative evaluation.

Following extensive investigation, Algorithm Audit has identified several ethical issues that it considers the most urgent and important. As part of the investigation, academic and domain experts and several stakeholders have been consulted. The results of this investigation form the basis for the deliberations of the advice commission, and can be found in the problem statement 'Predicting irresponsible driving behavior' (ALGO:AA:2025:01:P).

Additionally, Algorithm Audit has conducted a focus group involving users of shared mobility platforms. This case marks the first time Algorithm Audit has organized a focus group as part of its investigation, exploring how to best incorporate user perspectives into the evaluation process. The focus group results were submitted

in advance to the advice commission as general input to inform their deliberations, rather than as binding guidelines. The complete focus group findings are available in a separate document with focus group results. This advice report crystallizes the deliberative evaluation of a group of experts and stakeholders that together formed the advice commission. The commission is a diverse group in which various relevant disciplines and stakeholders are represented. The exact composition of the commission can be found under the section 7. Composition of the advice commission. Both the commission and Algorithm Audit have conducted this study independently from the car sharing platform. Neither the investigation nor the advice have been commissioned or funded by the platform. The advice of the commission, though non-binding, serves as a normative guideline for all parties that struggle with the responsible use of algorithmic risk profiling in the context of (car) sharing platforms.

2. Scope of advice

Regarding the case at hand, Algorithm Audit and the normative advice commission have identified key ethical issues that require normative assessment, particularly where existing regulations and guidelines do not provide ready-made answers. The specific issues addressed in this advice are:

1. **Purpose and validity of prediction:** Determining the legitimate purpose of the algorithm, the validity of the current model for achieving that purpose, and assessing when appropriate monitoring turns into inappropriate surveillance or paternalism;
2. **Transparency and explainability requirements:** Assessing what constitutes meaningful explanations to users about monitoring of their driving behavior;
3. **Variable selection for profiling:** Evaluating which driving characteristics may be justifiably used as input variables for predicting irresponsible driving behavior, and under what conditions;
4. **Model calibration and balancing false positives/false negatives:** Determining the appropriate balance between false positive and false negative predictions, considering the impacts of both.

That the commission concentrates on these issues does not mean they exhaustively cover the space of responsible use of risk profiling algorithms by the car sharing platform. For instance, general data quality control or aspects of good governance with regard to data processing, documentation, decision-making processes, and allocation of roles and responsibilities do not fall under the scope of this advice – not because they would be less important, but because existing frameworks already provide significant guidance for these

Box 1

Algoprudence: Case-based normative advice for ethical algorithms

Algorithm Audit does not have a mandate to issue legally binding rulings or official judgements. In our case studies, we give non-binding ethical advice. Ethical advice often goes beyond advice on what is required for legal compliance. Yet in the absence of legal rulings or clear standards established by a supervisory body, our independent ethical advice also serves as a preliminary signpost for organizations. Our case advice may also help elaborate official standards or support future decisions by legal bodies. In this sense, our ethical advice does have relevance for the legal domain.

aspects. An exception in this regard is the issue of model validity discussed at the outset, which is less an open normative question than a clear operational standard. Nonetheless, the commission has found it important to draw attention to this issue in the advice because of its fundamental importance to the rest of the case.
The scope of the advice document does not fully map onto the questions raised in the problem statement. This applies in particular to question 3 on sensitivity testing methods, through which the impact of hyperparameter selection on the accuracy of model predictions is examined, which has been left out of the advice. The reason is that the commission, instead of an urgent normative question, has determined it a matter of best practice that can be left to the professional judgement of data scientists.

3. General considerations and purpose of prediction

In determining how algorithmic risk prediction can be justifiably applied to identify irresponsible driving behavior in a car sharing platform's database, it is important to look at what algorithmic-driven profiling aims to achieve. For the current case, the commission identifies a significant concern regarding what the risk prediction model is actually predicting. The model has initially been trained on users who have been banned from the platform for various reasons, including not only driving behavior but also payment issues, late returns, leaving cars dirty, and other forms of misuse. The commission notes that there is a mismatch between the target variable in training (banned users for various reasons) and the objective of the model (irresponsible driving).[1]

The commission judges that this mismatch undermines the validity of the model and must be addressed. If the aim is to predict and reduce damage costs through unsafe driving, then the model should be specifically trained on cases where damage was caused by driving behavior, rather than on the broader category of banned users. The commission emphasizes that this issue must be resolved before the model can be considered valid for its stated purpose. The commission advises that proper implementation requires:

> Clear definition of what constitutes "irresponsible driving" specifically related to damage;
> Training data that specifically connects driving behaviors to actual damage cases;
> Features in the model that have a demonstrated relationship to the defined risk.

The commission has proceeded with its assessment based on this future improved state of the risk prediction algorithm, rather than the current implementation. Hence, all further advice in this report is issued by the commission under the condition that the current model will be improved to specifically predict damage-causing driving behavior rather than general misuse. This requires careful cleaning of training data and feature engineering, and potentially additional data collection to examine whether statistical relationships exist between driving behaviors and damage.

The commission raises the question at what point monitoring of driving behavior becomes excessive monitoring which goes beyond the core function of a car sharing service, which is primarily to provide access to vehicles. If the platform positions itself as an authority on driving behavior, and users feel constantly monitored and judged for how they drive, it creates a surveillance environment that is both suspicious and paternalistic

[1] The platform has indicated that at the time of the commission meeting, they had already flagged this issue. For the current advice, though, the commission has based its considerations on the model as it has been described in the problem statement.

towards users. Some commission members express the opinion that providing feedback on driving behavior, for instance nudging users towards safer driving, could be a service to users. Yet the commission generally emphasizes that this must be done in a way that respects user autonomy. The commission notes the difference between monitoring to reduce damage costs (a legitimate business interest) and imposing normative judgments about how users should drive beyond what's necessary for that purpose.

The commission recognizes that some driving behaviors, particularly speeding, have clear safety implications, while others like acceleration might depend on the context – sometimes representing risky behavior but other times simply reflecting personal driving style in safe conditions. The platform should be careful not to overreach in its behavioral guidance, particularly for behaviors that have an ambiguous or unproven relationship to safety or damage.

On the issue whether the current algorithmic approach is necessary at all, the commission has first of all considered the business rationale. Given that damage costs constitute approximately 7% (€2M-€3M) of annual revenue (€25M-€45M), the commission generally recognizes the legitimate business interest in reducing damage costs. The platform has indicated there has been some reduction in damage costs since implementation of the model, though they have not provided specific figures. Given the lack of demonstrated effects of implementing the machine learning-driven risk prediction model and insufficient consideration of alternative methods, the commission has withheld judgment on whether the current algorithmic approach is proportionate and effective for this purpose compared to alternatives. It has suggested simpler and more transparent approaches might achieve similar goals, such as:

> Rule-based profiling with clear, manually set thresholds;
> Simpler warning systems for specific behaviors (e.g., direct warnings after specific incidents rather than based on machine learning-driven risk prediction);
> Alternative monitoring methods, such as having users inspect cars before and after use to document the condition of the car.

The commission recommends that the platform thoroughly evaluate these alternatives before proceeding with a complex algorithmic system.

4. Transparency and explainability

When using risk profiling methods to identify potentially irresponsible driving behavior, it is necessary that users are aware that driving behavior is monitored and that decisions can be explained to users. This is an important principle for maintaining legitimacy and trust in the platform.

Subscription to the platform's services should be conditioned on user consent, not just regarding data collection, but also regarding the monitoring of individual driving behaviour. The commission emphasizes that user consent should qualify as informed consent. The commission expresses concern that consent processes, where users accept a lengthy privacy statement, may lead to formal agreement without genuine awareness that individual driving behaviour is monitored.[2] The commission considers it reasonable that users unwilling

[2] To prevent unintentional consent, the platform also uses pop-up notifications within its application to inform users about the monitoring of driving behavior.

to have their driving behavior data collected can choose to unsubscribe from the platform's services after receiving the notification.
When users receive a warning about their driving behavior, the explanation should be specific enough to help them understand which behaviors are problematic. Current generic notifications about "irresponsible driving" without further details are insufficient. The commission advises that warnings include specific driving behaviors that triggered the warning context for why these behaviors are considered problematic, and clear guidance on how to improve. While detailed breakdowns of all data points might overwhelm users, the most significant factors contributing to the warning should be clearly communicated, so that the user for instance knows it is hard braking or speeding in this or that recent trip that helped trigger the warning.

The commission does not deem the current list of aggregated features to be sufficient for such kind of meaningful explanation. They for instance include a lot of overlap between the variables that capture speeding behaviour in a different way. These variables overall represent highly technical, complicated categories (such as 'total driving events'), that are not directly insightful to the average user (see Box 2). Simply providing the user with the feature(s) from this list that have contributed most to the risk score will not constitute a meaningful explanation. The commission suggests using data processing methods to come up with more useful categories that can be communicated to users. More detailed, technical information (i.e., directly relating to the original features) can and should always be provided to the user on request.

The commission notes that the style of communication significantly impacts how users perceive the platform. Testimony from users who have received a warning indicates that the current communication approach of the platform comes across as overly strict and suspicious, which creates anxiety and may discourage users from further using the platform. Because users are currently often unaware their driving behavior is being

Box 2

Translate variables in understandable features

Not meaningful explaination

During your last trips we have observed high:

TOTAL_DRIVING_EVENTS_PERKM

These behaviors lead to an increased risk of damages to our vehicles. We kindly ask for your attention to these matters to help prevent damages.

More meaningful explaination

During your last trips we have observed your behavior and noticed:

> Excessive hard breaking;
> Taking corners at excessive speed.

These behaviors lead to an increased risk of damages to our vehicles. We kindly ask for your attention to these matters to help prevent damages.

TOTAL_DRIVING_EVENTS_PERKM is a combination of the variables TOTAL_CORNERING_EVENTS_PERKM and TOTAL_BRAKING_EVENTS_PERKM

monitored[3], sudden warnings about irresponsible behaviour feel invasive, especially if it comes across as if you are directly suspected of wrongdoing and potentially blocked, even though driving behaviors such as hard braking or cornering may sometimes have legitimate explanations. Instead, the platform's communication should adopt a cooperative rather than an adversarial tone, emphasizing the shared interest in safe driving and vehicle maintenance, nudging the user towards responsible driving behavior.

The commission avoids declaring what precisely would be the most meaningful form of explanation, given that this must be asked to users themselves. The commission suggests testing out alternatives and getting feedback on them from actual users. The concern about paternalistic effects discussed above should be also taken into consideration when asking feedback from users about the right kind of communication about driving behavior.

A first goal of transparently communicating why a warning has been sent, also for legal reasons, is the possibility of redress. The commission states the platform should establish and clearly communicate mechanisms for users to challenge decisions they believe are factually incorrect or unjustified. These mechanisms should be easily accessible and responsive. The commission appreciates that human analysts currently review algorithmic recommendations before warnings or blocking decisions are issued. Users should be informed that algorithmic risk scores are not the sole determinant of warnings or blocking, that human review takes place before actions are taken, and that there are clear avenues for appeal or providing additional context.
Another aim of transparency is to give users more general insight into their driving behavior. While the commission warns against the paternalistic effects of giving continuous feedback to users on how they drive, commission members have recognized that some users may want to access such information. The commission suggests exploring possibilities of creating an online dashboard where users can learn about their driving metrics over time. For those users who have received a warning earlier, it might give them reassurance to see their driving has indeed improved and they are no longer at risk of being blocked. It may increase transparency about what data is being collected, and build trust through openness about the monitoring process.

5. Variable selection for profiling

Presupposing that the adjustments outlined in section 1 regarding the validity of the model are made, the commission does appreciate that only driving data is used for risk prediction, rather than other personal data such as age, ZIP code, or years holding a driving license. The commission also positively acknowledges that the platform works with aggregated data (e.g., speeding events per kilometer) rather than seeing exactly how much a user was speeding in specific instances, which provides some privacy protection for users.

Nevertheless, careful consideration should be given to which driving behavior variables are justified for inclusion in the risk prediction model. The commission recommends, among other considerations, to minimize the number of variables used, to only include variables which have a clear and direct connection between driving behaviour and damage and to consider how driving behaviour might correlate to protected demographic characteristics, e.g., ethnicity and gender, in order to check if the model does not perform

[3] Note that users are informed at the beginning of every trip about data collection through pop-up notification in the app of the platform. Nonetheless, user experience still tells us that this may not adequately raise active awareness.

differently for various groups in an unjustified way (see below).

Among the driving behavior features, the commission considers speeding as the most legitimate variable to include in the model. There is a clear and direct connection between speeding and unsafe driving behavior that may lead to accidents and damage. The commission generally agrees that speeding is an objectively problematic behavior that users have control over, marking it a reasonable basis for risk assessment.

Regarding speeding data, the commission draws attention to reliability problems that should be addressed. As the platform indicates, there have been instances where GPS inaccuracies have resulted in incorrect speed limit assessments. The platform has attempted to mitigate this by filtering out implausible speeding events, but this remains an area of concern. The commission recommends investigating further refinement of GPS accuracy, improved filtering of implausible speeding data, and other methods to reduce false positives. Improvements in transparency, explainability and redress mechanisms as outlined in section 2 have to ensure that users can easily flag incorrect speeding measurements.

The commission has expressed that other variables such as fast acceleration, heavy braking, and hard cornering require more consideration than speeding. These behaviors may not always indicate irresponsible driving and may be influenced by contextual factors beyond the driver's control. The commission notes that hard acceleration might be harmless in certain contexts, such as accelerating at a green light when no other vehicles are present. Similarly, hard braking may indicate a responsible safety intervention rather than irresponsible driving – the driver is responding appropriately to an unexpected situation.

The commission also observes that driving environments significantly impact these variables. Driving trips in urban areas and during rush hour will naturally include more sudden and hard brakes than those in rural areas, regardless of how responsibly users drive. This raises concerns about potential adverse effects for specific demographic groups, as the model might disadvantage drivers who primarily use vehicles in urban areas or specific neighborhoods.

While the commission does not recommend categorically excluding these variables, it advises that their use must be clearly justified with evidence showing a direct connection to damage risk. The platform should conduct further analysis, e.g., disaggregated data analysis, to verify that these variables are not serving as proxies for (sensitive) geographic or demographic factors and to check if the model does not perform differently for various groups in an unjustified way. The commission also recommends consulting representative focus groups that include diverse demographics, particularly people from marginalized communities and those with medical conditions. Certain driving behaviors flagged as risky – such as hard braking – could be related to medical conditions rather than irresponsible driving. These diverse user groups should be involved in the development process of the algorithmic-driven risk assessment process, helping to decide which variables are included and how.

The commission strongly recommends applying the principle of data minimization – using only the variables that are demonstrably necessary and proportionate for the legitimate purpose of predicting damage risk. The platform should conduct and document a thorough feature selection process, demonstrating which variables contribute significantly to prediction accuracy and which could be removed without substantial impact. This does not only reduce unfair penalization of driving behavior that does not significantly increase the risk of

damage. It may also improve the model's transparency and explainability.

6. Model calibration, FP/FN-balancing

Evaluating a risk prediction model, there is an important value-dimension to how the model is calibrated, in particular with regard to the trade-offs between false positive and false negative rates. Usually these rates are interdependent, where a relative decrease in false positives (so an increase in true positives) at the same time means an increase in false negatives – as illustrated in Figure 1. To evaluate the way these should be balanced, the commission first considers the impacts of both types of errors: false positives (responsible drivers incorrectly classified as irresponsible) and false negatives (irresponsible drivers incorrectly classified as responsible). See also Box 3.

The committee notes that minimizing the false negatives (FNs) aligns with the primary objective of the risk prediction model. In evaluating these risks, safety and business considerations must be weighed against user experience and trust. While the commission expresses concern about both FPs and FNs, there is general agreement that the presence of a human review process mitigates some of the risks associated with FPs. That is to say, meaningful human review can ensure that a substantial portion of FPs resulting from the model is caught before warning messages are unduly sent to users who drive responsibly.

The commission states that without improvement on the model validity, it is impossible to adequately evaluate the concrete scenarios given in the problem statement (Figure 1) with various true positive (TP) and FP rates. This issue notwithstanding, the commission makes some suggestions. First, when considering rates, it is more helpful to translate these abstract metrics into more tangible user experiences: a FP rate = 0.0157 means approximately 1 out of 64 rides would result in a FP, meaning a frequent user might be incorrectly flagged by

Box 3

Interpretation of false positives and false negatives

Concerns regarding false negatives (FNs):

> Material risks for the platform from not flagging users who are more likely to cause vehicle damage, resulting in increased damage costs;
> Reduced road safety for all users and the general public;
> The platform's vehicles potentially becoming associated with irresponsible driving behavior on the road.

Concerns regarding false positives (FPs):

> Users experiencing false accusations of irresponsible driving behavior;
> Stronger feeling of being under surveillance;
> Experience of the platform being unfair;
> Risk of users switching to competing platforms due to negative experiences, incl. loss of reputation of the platform;
> Material costs for employing more human analysts to examine FPs;
> Certain groups potentially being disproportionately subject to false positives, introducing bias and disparate impact;
> Users not taking warnings seriously if too many false warnings are issued.

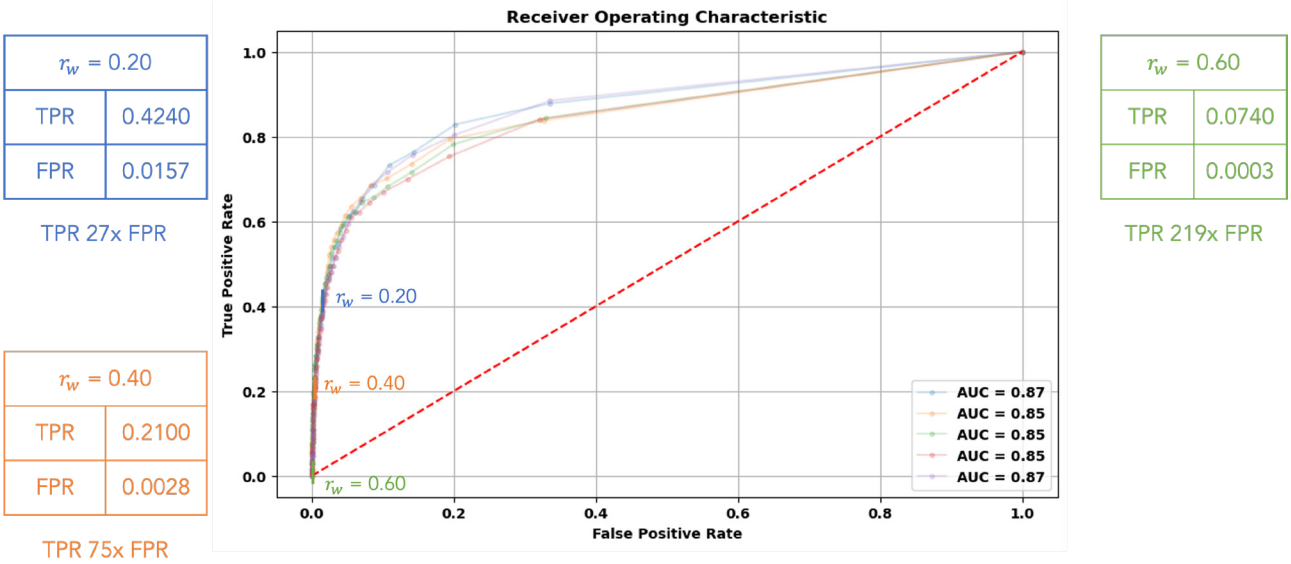

Figure 1 - ROC curve of risk Balanced Random Forest (BRF)-prediction model for 5-fold cross-validation

the model roughly once every two months of daily use. Under this framework, the commission provisionally suggests that one unwarranted indication per year constitutes an acceptable threshold for frequent users. Further research among users should provide more grounded guidelines for what users generally find acceptable. The commission notes that this assessment is contingent on maintaining an effective human review process and ensuring the fairness of the system across all user groups.

Second, if we do take the above scenarios as reference point (even though they will change when addressing the model validity), the commission observes that in this case the FP rate is generally low. In the current process human analysts review all cases flagged by the model, and in approximately 50-60% of the cases the analysts override the model's recommendation. The commission sees this as a positive sign that meaningful human discretion is being exercised, suggesting that the human review represents a meaningful evaluation rather than merely rubber-stamping the model's outputs. It is a further indication that the risks of FPs resulting from the model are relatively low and mitigated already, meaning there is more room for focusing on decreasing and mitigating FNs. The commission does recommend performing additional analyses to evaluate whether particular groups have a higher FP rate than others. Aside from a clear business interest of preventing damage costs, it is in the interest of general safety to flag truly irresponsible driving behavior.

These recommendations are made under the condition that this substantial human element in the review process is maintained, ensuring analysts have sufficient information and training to make fair assessments, and considering ways to document decision rationales for accountability, consistency, and continuous improvement. Moreover, the recommendations outlined in section 2 are needed to mitigate the risks of FPs, adopting for the warnings a more constructive, helpful and less suspicious tone towards users. In this way, receiving a warning is not as impactful, even when the user is a responsible driver.

7. Composition of the advice commission

This advice is the outcome of a common deliberative process. Hence, specific claims in this document do not necessarily align with the opinion of individual members of the audit commission. Members of the commission cannot be individually held responsible for this advice.

Date

The audit commission met physically in The Hague on 13 January 2025. This advisory document has been approved by all members of the advisory commission on 26 June 2025.

Advice commission members

The normative advice commission for this case is formed by:

> Cynthia Liem, Associate Professor at the Multimedia Computing Group, TU Delft
> Hilde Weerts, Assistant Professor Fair and Explainable Machine Learning, TU Eindhoven
> Joris Krijger, AI & Ethics Officer, De Volksbank
> Maaike Harbers, Professor of Applied Sciences (lector) Artificial Intelligence & Society, Rotterdam University of Applied Sciences
> Monique Steijns, Founder The People's AI agency
> Anne Rijlaarsdam, user car sharing platform.

A data scientist representing the car sharing platform was also present during the commission meeting to answer factual questions, but this person is not part of the advice commission.

Acknowledgements

Besides many individuals we have spoken to or those who have diligently read our work, we extend special gratitude to the following people and organizations for their valuable contributions to this project:

> Vardâyani Djwalapersad
> Tom Driessen
> Joel Persson

About Algorithm Audit

Algorithm Audit is a European knowledge platform for AI bias testing and normative AI standards. The goals of the NGO are four-fold:

Knowledge platform

Bringing together experts and knowledge to foster the collective learning process on the responsible use of algorithms, see our white papers and public standards.

Normative advice commissions

Advising on ethical issues that arise in concrete algorithmic practice through deliberative and diverse normative advice commissions, resulting in algoprudence

Technical tools

Implementing and testing technical tools to detect and mitigate bias, e.g., sociotechnical evaluation of generative AI, unsupervised bias detection and synthetic data generation.

Project work

Supporting public and private sector organisations with specific questions regarding responsible use of AI, from a not-for-profit perspective.

Structural partners of Algorithm Audit

SIDN Fund

The SIDN Fund stands for a strong internet for all. The Fund invests in bold projects with added societal value that contribute to a strong internet, strong internet users, or that focus on the internet's significance for public values and society.

European AI&Society Fund

The European AI&Society Fund supports organisations from entire Europe that shape human and society centered AI policy. The Fund is a collaboration of 14 European and American philanthropic organisations.

Building *AI auditing* capacity from a *not-for-profit* perspective

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Parkstraat 22, 2514 JK The Hague

Algorithm Audit

Stichting Algorithm Audit is registered as a non-profit organisation at the Dutch Chamber of Commerce under license number 83979212