Fade to grey: from black-box to AI-antagonistic audits

ERWAN LE MERRER, is a Researcher at *Univ Rennes, Inria, CNRS, IRISA*, Rennes, France
GILLES TREDAN, is a Research Director at *LAAS-CNRS*, Toulouse, France

We argue for a change of focus in AI auditing: from assuming truthful model responses to considering the antagonistic interests of an auditor and the AI model under scrutiny. This change must come with increased and enforceable assumptions on the nature of the AI being audited.

Opinion. Submitted: 21/10/24. Accepted: 04/12/25. Publication: 08/26 (scheduled).

Additional Key Words and Phrases: AI models, audits, fairness

ACM Reference Format:
Erwan Le Merrer and Gilles Tredan. 2026. Fade to grey: from black-box to AI-antagonistic audits. 1, 1 (April 2026), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

In the classic movie Blade Runner, Inspector Deckard is performing a Turing test (coined *Voight-Kampff* in the movie) on Rachael, by asking her questions to unmask her potential robotic ("replicant") nature. Rachael, as we will learn later in the movie, does not know that she is a replicant, and thus appears to respond truthfully to Deckard's questions.

This is the current status of assumptions in the AI or algorithm auditing research field. Researchers are willing to provide tools for assessing the compliance of AI models concerning the lack of bias or discrimination, for instance. Today's critical models are executed at third parties (termed Very Large Online Platforms in the European AI act for example), and then are not directly observable in a white-box scenario using explanation tools. Auditors have to interact with them using requests, and observe their resulting responses. The term black-box has emerged in this context [2]. In this setup, an auditor is supposed to have no prior knowledge about the model's internals, besides how to interact with it, and its general goal (e.g. tagging images, giving a go for a bank credit).

Black starts fading. Yet in practice, many academic works use the term black-box audit, while assuming some prior knowledge of what is inside the box. One of the most common example being that the test distribution on the audit input is drawn with the same law as the one used to train the black-box model. Another common assumption is the knowledge of the input space of the algorithm being audited: the

Authors' Contact Information: Erwan Le Merrer, erwan.le-merrer@inria.fr, is a Researcher at *Univ Rennes, Inria, CNRS, IRISA*, Rennes, France; Gilles Tredan, is a Research Director at *LAAS-CNRS*, Toulouse, France, gtredan@laas.fr.

ACM XXXX-XXXX/2026/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

YouTube recommender algorithm takes thousands of input features into account, of which in-field sock-puppet audits leverage only a fraction due to opacity.

The recent paper by Casper et al. [4], with the definitive title "Black-Box Access is Insufficient for Rigorous AI Audits", states that "Many black-box and grey-box attack algorithms are simply indirect or inefficient versions of white-box ones"; stressing the fact that more assumptions are required to perform audits. This seems correct, yet one must not forget that this restrictive black-box setup is more a constraint than a choice for the auditor operating when facing a third party. As in the case of Deckard (who cannot realistically ask thousands of questions to perform his Turing test), researchers prove as well that auditing black-box explanations of decisions from a model incurs an impractically large amount of queries [3].

In this context, providing the auditor with additional prior knowledge appears to be a natural solution to overcome the hardness of auditing. Alas, white-box approaches come at the cost of genericity. The beauty of black-box audit approaches is their ability to produce results for all boxes, regardless of their internals, whereas in the extreme case each white-box approach has to be specifically tailored to its target to enable the efficiency sought after. Given the pace at which platforms evolve, this might reveal equally prohibitively costly. Moreover, as the auditor increases the set of assumptions leveraged for her audit, she also increases the manipulation opportunities by the platform. In other words, she increases her audit attack surface.

The overlooked latent antagonism in auditing. Indeed, and to make things even more complicated for auditors, measurement techniques by the research community have vastly overlooked a core antagonism at play. On the one hand, an auditor wants to perform an accurate assessment of a model; on the other, the operator of this model wants to optimize its accuracy (to maximize its profits for instance). And these two are often conflicting, as notoriously illustrated in the engineering realm by the "Diesel gate" scandal, where a company had implanted a deceptive mechanism in cars so that carbon emissions are lower when the car detects a measurement setup (i.e. it is being audited). In the research community, this problem has been identified as "fairwashing" [1], a practice in which model operators would manipulate the audit to pass a fairness assessment test. Interestingly, this is also a Blade Runner scene where the replicant named Léon is actively trying to cheat the Turing test, by responding as "humanly" as possible, before finally being uncovered.

First attempts to operate in this antagonistic setup end up with negative results. Fukuchi et al. [3] reason on how a platform can cheat in providing a supposedly representative sub-dataset, so that fairness tests on it are validated, while the distribution of this cheated dataset is indistinguishable from a legitimate one for the auditor. In other words, the platform can always provide a cheated dataset, and this dataset complies with what is expected from a fair dataset. A malicious platform thus can lie just below the level of potential detection, using precise measures like the Wasserstein distance between legitimate and manipulated data points in the dataset in order to fairwash its operation. This result is demonstrated in the setup where the auditor knows the actual distribution of the data at the platform (i.e. non black-box), which means that this is obviously harder when one auditor does not have this assumption.

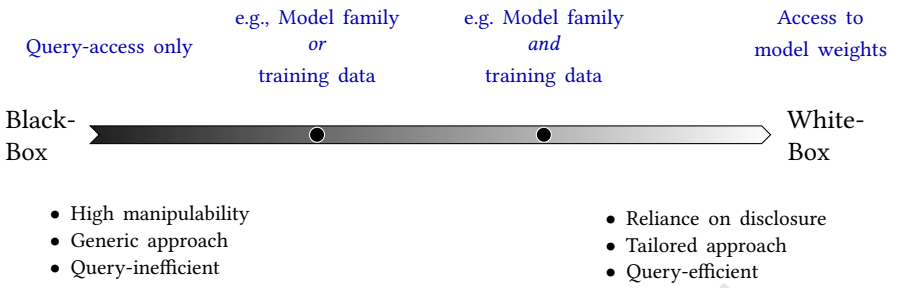

Fig. 1. Examples of assumptions and their consequences on the outcome of AI-model audits.

A second thread concerns manipulation-proofness [7]: the auditor will actively attempt to avoid being gamed by the platform. Auditors want to enforce that after an initial audit with a given query set for the estimation of fairness, the platform cannot change its model to an unfair one (that is, post-audit), while staying consistent with its previous answers. To that end, the assumption made by this work is that the platform declares the hypothesis class of its current model; a hypothesis class being for example all the possible instances of a random forest algorithm, or a fully connected neural network with three layers. In this interesting setup, estimation guarantees are possible. This constitutes a first rigorous step to a "grey"-box setup, acknowledging a departure from the black-box setup for having stronger guarantees in front of a potentially manipulative platform.

Yet, despite being more restrictive for the platform than a black-box setup, manipulation-proofness in practice cannot be achieved efficiently facing high-capacity models [5], possessing a large number of parameters and degrees of freedom to learn their task (and formally captured by the VC-dimension for instance). These models are, unfortunately for the auditor, those that are massively driving the current AI rush, and are able to "bend" to cope with the consistency of the limited query set by the auditor, while arranging the rest of their outputs on the metric they seek to optimize (e.g. accuracy instead of fairness).

These negative results come along with others that are dedicated to showing that explanations of decisions can also be manipulated without the possibility of detection [1, 6]. This is a clear motivation for designing robust audit algorithms (or antagonistic-audit algorithms), which are audit procedures that can cope with manipulation, in that they can estimate the target bias as correctly as possible, together with an objective of spotting actual manipulation attempts in a platform's outputs.

Conclusion: the quest for the "darker" assumption set. There is thus a twofold impediment to performing robust audits. The original black-box and truthful assumption audit setup is challenged today, as this setup is not suited for provable guarantees under a limited number of audit requests; on the other hand, the problem is aggravated because it is not realistic anymore to assume that the platform being audited will never attempt to game the audit.

This calls for a stronger set of assumptions on the model being audited, leading to a grey-box setup. Among the many possible shades of gray, one should weigh each

Fig. 2. Black-boxes that turn to grey (by Copilot).

hypothesis as a double-edged sword: each enables more efficiency for the auditor, but each specializes the approach and imposes a greater reliance on platform truthfulness and cooperation. In this balancing act, we believe one should seek assumptions on models that are "as dark as possible", as those are the best candidates to withstand both the rapid technological evolution of platforms and their potential antagonistic behavior.

It goes without saying that the information a platform accepts to disclose about its model, and that will serve as a work assumption for researchers and auditors, must come with solid means to enact their verification. Robust grey-box approaches are only as robust as the hypotheses they rely on. The validity of these hypotheses must be controlled by external means, such as enforcement coming from legal procedures to access data on servers.

References

[1] Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. (2019). Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*.

[2] Beizer, B. (1995). *Black-box testing: techniques for functional testing of software and systems*. John Wiley & Sons, Inc., USA.

[3] Bhattacharjee, R. and von Luxburg, U. (2024). Auditing local explanations is hard. In *Conference on Neural Information Processing Systems*.

[4] Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al. (2024). Black-box access is insufficient for rigorous ai audits. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.

[5] Godinot, A., Le Merrer, E., Trédan, G., Penzo, C., and Taïani, F. (2024). Under manipulations, are some ai models harder to audit? In *Conference on Secure and Trustworthy Machine Learning*.

[6] Le Merrer, E. and Trédan, G. (2020). Remote explainability faces the bouncer problem. *Nature machine intelligence*, 2(9):529–539.

[7] Yan, T. and Zhang, C. (2022). Active fairness auditing. In *International Conference on Machine Learning*.