ZEW policy brief

// NO.02 | FEBRUARY 2023

// Johannes Walter
(ZEW Mannheim & KIT)

Human Oversight Done Right: The AI Act Should Use Humans to Monitor AI Only When Effective

The EU's proposed Artificial Intelligence Act (AI Act) is meant to ensure safe AI systems in high-risk applications. The Act relies on human supervision of machine-learning algorithms, yet mounting evidence indicates that such oversight is not always reliable. In many cases, humans cannot accurately assess the quality of algorithmic recommendations, and thus fail to prevent harmful behaviour. This policy brief proposes three ways to solve the problem: First, Article 14 of the AI Act should be revised to acknowledge that humans often have difficulty assessing recommendations made by algorithms. Second, the suitability of human oversight for preventing harmful outcomes should be empirically tested for every high-risk application under consideration. Third, following Biermann et al. (2022), human decision-makers should receive feedback on past decisions to enable learning and improve future decisions.

KEY MESSAGES

- Humans often cannot accurately assess the quality of algorithmic advice and commonly fail to correct harmful AI decisions.
- The proposed AI Act, therefore, should not rely unquestioningly on humans to prevent harm and should instead require tests assessing the feasibility and efficacy of human oversight.
- The tests should also investigate whether providing feedback can facilitate and improve human oversight.
- If these tests find that human oversight fails to prevent harm, or even exacerbates harmful outcomes, then human oversight should not be relied upon.

MITIGATING HARM AS PROPOSED IN THE EU'S ARTIFICIAL INTELLIGENCE ACT

Increasingly, AI systems are supporting human decision-making. In healthcare, algorithms recommend which patients should undergo expensive treatments; in the justice system, they predict a defendant's risk of recommitting a crime; and in hiring decisions, they suggest which job applicants to invite for an interview. The Artificial Intelligence Act (AI Act) proposed by the European Commission would regulate the application of decision-supporting algorithms, among other aspects of AI.

The AI Act envisions safe, AI-supported decisions based on human oversight.

The AI Act would introduce a risk-based classification: low-risk applications would not be regulated, while the highest-risk applications, such as social credit scoring or biometric recognition, would be prohibited. In between these two ends of the risk spectrum are applications that could cause harm to health, safety or fundamental rights. These high-risk applications would be subjected to compulsory conformity assessments (Article 19) and human oversight (Article 14). The conformity assessments test algorithmic systems prior to and in some cases during deployment. Human oversight refers to the idea that a natural person can oversee and, when necessary, overrule the algorithmically derived recommendations.

So far, debates about the AI Act have focused on the definition of AI, the types of AI to be prohibited, and how to preserve incentives for innovation. This policy brief considers the problems associated with human oversight of AI, which have not yet received the attention they deserve. In brief: humans are not always reliable monitors of AI. Accordingly, they should be used only when their efficacy in preventing harmful outcomes is proven for a given case.

THE SHORTCOMINGS OF HUMAN OVERSIGHT

The possibility of human error in the supervision of AI receives no consideration in the AI Act. Recent studies have shown this to be a weighty omission. Biermann et al. (2022) found that participants in an online laboratory experiment were unable to identify and improve incorrect algorithmic recommendations. Likewise, growing evidence indicates that in many settings, humans are incapable of accurately assessing the quality of an advising algorithm and perform poorly when deciding whether to correct a suggestion (Green, 2022; Lai & Tan, 2019; Springer et al., 2017; Goodwin & Fildes, 1999). One potential explanation for this phenomenon is automation bias, a well-documented phenomenon in which humans tend to prefer the suggestions of algorithmic systems and to disregard contradictory information not produced through automation, even when it is accurate (Parasuraman & Manzey, 2010; Skitka et al., 1999).

Human oversight can be unreliable in many situations

Remarkably, humans make two general types of error: failing to correct wrong algorithmic advice and falsely correcting right algorithmic advice. Both errors have been shown to occur in real-world applications. For example, Fussey and Murray (2020) found that police in the United Kingdom severely overestimated the accuracy of a facial recognition system, assuming that it could correctly identify suspects at three times its actual rate. And Hill (2020) describes the first known case in the United States of a man being wrongfully arrested due to misidentification by facial recognition software. Evidence for the second type of error emerged in a study by Human Rights Watch (2017), which determined that judges across different US jurisdictions commonly overrule algorithmic recommendations in ways that harm defendants. In a later study, Stevenson & Doleac (2022) showed that if the judges had adhered to algorithmic recommendations, incarceration rates would have been considerably lower without loss to public safety.

CAN HUMAN OVERSIGHT BE IMPROVED?

Given its many problems, is human oversight doomed to fail, or can its efficacy be improved? Biermann et al. (2022) tested two interventions, one giving human decision-makers an explanation of the advising algorithm and the other giving them feedback on past decisions. When provided with a simple explanation of a biased advising algorithm, more study participants could identify the bias than could without additional information. But contrary to what one might expect, the average decision quality of participants did not improve, as it turns out that the ability to identify poor recommendations and the ability to correct them are separate skills. This finding is in line with the literature on explainability of AI: Alufaisan et al. (2021) draw the conclusion that an explanation alone is no panacea; whether it improves a decision depends on various factors.

The situation is different in the case of feedback. Biermann et al. (2022) found that feedback helped participants to improve their performance as well as their ability to identify bias. Hence, it appears that feedback allows humans involved in oversight to learn about the algorithm and the reliability of their own decisions. Also, feedback about past outcomes provides a reference point. In view of the wide variety of AI applications, providing feedback will not always be possible or helpful. Yet in a stable decision environment – one in which algorithm, task, data generation and human decision-maker remain the same – it seems like a promising approach.

Human decision makers can receive feedback to improve future decisions.

RECOMMENDATIONS

Since human oversight of AI can be unreliable, some researchers have suggested that it be abandoned altogether (Green, 2022). Yet this policy brief argues that human oversight does have value in preventing harm. For example, De-Arteaga et al. (2020) found in a real-world context that human decision-makers were able to overrule many erroneous algorithmic recommendations. In practice, the feasibility and efficacy of human oversight depend on a variety of factors, including the context, the AI, the level and type of insight and the expertise of the human decision-makers. Accordingly, human oversight should not simply be jettisoned. What it needs is more careful implementation. To this end, we propose the following measures:

The feasibility and efficacy of human oversight should be empirically tested for every high-risk use case under consideration.

- **First**, the AI Act should acknowledge that human oversight is not always reliable. Article 14 could thus be amended to include a clause about the risks of human oversight.
- **Second**, tests assessing the feasibility and efficacy of human oversight in preventing harm should be mandatory for high-risk AI applications. The test requirements could be included under the conformity assessments in Article 19 of the proposed AI Act. Specifically, the tests should determine whether humans are capable of accurately assessing the quality of algorithmic advice and whether they are able to overrule it when necessary. In their most elementary form, such tests would compare actual outcomes under human oversight with the hypothetical outcomes that would have resulted without human intervention.
- **Third**, when feasible and appropriate, these tests should include information intended to improve the human decision-makers' ability to assess and correct algorithmic recommendations. To that end, feedback about the outcomes of past decisions seems to be a promising intervention.
- **Finally**, if the tests reveal that human oversight does not prevent harm, that it exacerbates harmful decisions or that it introduces new types of bias, it should not be relied upon. In these cases, the AI Act should include a mechanism that limits or prohibits human oversight.

To enable learning and improve oversight, human decision-makers should receive feedback on the outcome of past decisions.

REFERENCES

Alufaisan, Y., Marusich, L. R., Bakdash, J.Z., Zhou, Y. & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35. 6618 – 6626.

Biermann, J., Horton J. J. & Walter, J. D. (2022). Algorithmic advice as a credence good. ZEW Discussion Paper No.22-071.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

European Commission (2021). Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).

Fussey, P. & Murray, D. (2020). Policing uses of live facial recognition in the United Kingdom. Regulating Biometrics: Global Approaches and Urgent Questions. AI Now Institute.

Goodwin, P. & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy? Journal of Behavioral Decision Making.

Green, B. (2022). "The Flaws of Policies Requiring Human Oversight of Government Algorithms," Computer Law & Security Review, vol. 45.

Hill, K. (2020). Wrongfully accused by an algorithm. The New York Times.

Human Rights Watch (2017). "Not in it for Justice": How California's Pretrial Detention and Bail System Unfairly Punishes Poor People.

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.

Parasuraman, R. & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors.

Skitka, L. J., Mosier, K. L. & Burdick, M. Does automation bias decision-making? International Journal of Human-Computer Studies.

Springer, A., Hollis, V. & Whittaker, S. (2017). Dice in the black box: User experiences with an inscrutable algorithm. The AAAI 2017 Spring Symposium on Designing the User Experience of Machine Learning Systems.

Stevenson, M. T. & Doleac, J. L. (2022). Algorithmic Risk Assessment in the Hands of Humans.

ZEW policy brief

Author: Johannes Walter (ZEW Mannheim & KIT) · johannes.walter@zew.de

Publisher: ZEW – Leibniz Centre for European Economic Research
L 7, 1 · 68161 Mannheim · Germany · info@zew.de · www.zew.de/en · twitter.com/ZEW_en
President: Prof. Achim Wambach, PhD · Managing Director: Thomas Kohl

Editorial responsibility: Dr. Frank Herkenhoff · kommunikation@zew.de