9th International Conference on Computational Social Science IC$^2$S$^2$
July 17-20, 2023, Copenhagen, Denmark

# Blame attribution in human-AI and human-only systems: Crowdsourcing judgments from Twitter

*Keywords: Blame; Attribution; Artificial Intelligence; Human-AI systems; Twitter*

This paper proposes and tests a method to study and compare people's blame attributions in human-AI and human-only systems as publicly stated on online platforms. Incidents involving AI are a contemporary topic that is often talked about on Twitter. Previous research studying attribution towards autonomous artificial agents has mostly used vignettes (Franklin, Awad, & Lagnado, 2021). The present paper proposes a more ecologically valid method that may replicate and contextualize previous experimental findings, or identify new factors that are relevant to people's blame attributions. Specifically, the goal of the method is to identify the *agents* people are attributing blame to, and the *factors* that have an effect on people's attributions. Agents are often context-specific, thus requiring a bottom-up approach to identifying which ones are relevant. The investigated factors that are highly correlated with blame build on the framework proposed by Franklin, Ashton, Awad, and Lagnado, examining how these factors are used by people making attributions outside the context of an experimental study. The nature of online data - written, real-world instances of human behaviour - makes it possible to capture naturally-occurring behaviour that reacts to critical events, rather than behaviour in an experimental setting. It is open-ended, that is "participants" are not assigned a task, thus allowing researchers to conduct more hypothesis-free. exploitative studies (Kapoor et al., 2018).

We collected relevant tweets using Twitter's official API. We focused our sample on tweets responding to famous incidents involving AI and their human-only equivalents. Specifically, the study looked at tweets responding to the Ofqual A-levels predictive algorithm (i.e., a computer program designed to predict what grades the students would have received if they had taken exams), the COMPAS recidivism algorithm (i.e., algorithm designed to assess the risk that a given defendant will commit a crime after release), the self-driving Uber hitting and killing someone, the self-driving Tesla crash, the Amazon hiring algorithm scandal (i.e., an algorithm that would assess people's job application that had a bias against female applicants), and the use of AI-generated art and text. We also collected tweets on similar incidents involving human-only systems, where no AI was involved, to enable a comparison of the prevalence of factors and sentiment across the two contexts; we collected tweets on road accidents involving human drivers and university admission scandals (e.g. operation varsity blues).

We followed a hybrid qualitative-quantitative approach to explore blame allocation amongst agents and factors in a bottom-up manner in two stages. The first stage in that process involved manual qualitative coding of the tweets in terms of *blame attribution, agents, factors, and sentiment*, applying a variation of the framework method (Ruhl, 2004). The second stage involved transforming the codes created in the first stage into variables, investigating potential associations between them, and applying an unsupervised classification algorithm to examine how these variables cluster together. Five trained coders independently coded a sample of 1342 tweets for whether they involved a responsibility attribution for the pre-specified contexts. The subset of tweets that involved an attribution (N=522) was then coded in terms of the agents (e.g., the developer) and factors (e.g., capability), and the sentiment of the attribution, which was either positive, negative, or neutral. 28% (N=465) of tweets were blindly double-coded.

We used two quantitative analysis methods to explore this dataset based on the first-stage qualitative analysis. We used Pearson's r statistic to test for associations between blame, context, sentiment, agents, and factors in the subset of tweets that involved an attribution. Using the coding of agents, factors, sentiment, and blame attributions, we applied a k-means clustering to the dataset comprising all corpora to identify clusters and confirm whether they predict blame attribution. We identified 10 agents and 12 factors related to attribution across the different contexts. Agents with a high proportion of blame attributions were the algorithm (28%), company (13%), government (12%), and system (11%). Incidents involving human-only systems had a higher prevalence of factors relevant to moral attributions, such as bias (57%), obligation (67%), and intent or foreseeability (67%). On the flip side, incidents involving AI had a higher prevalence of factors relating to performance and use, such as capability (81%) negative result (100%), and replacement (100%). Blame was most strongly positively correlated with human-only scenarios ($r(703) = .22, p = <.001$), negative sentiment ($r(703) = -.72, p = <.001$), and bias ($r(703) = .21, p = <.001$). The k-means cluster analysis grouped the variables into 6 clusters. Clusters 3, 5, and 6 grouped together tweets containing >94% blame attributions.

This study replicated previous findings that humans get more blamed for their intent or foresight, and algorithms get blamed more for their role or capability mirror the present results (Franklin et al., 2022). Our finding that obligation is more prevalent for human-only systems, is in line with previous evidence showing people are more likely to blame someone up the chain of command when a machine makes a mistake (Hidalgo, Orghian, Canals, De Almeida, & Martin, 2021). Negative outcomes are more prevalent in human-AI contexts is in line with the finding that machines get more blame for their outcomes (Hidalgo et al., 2021). Given the novelty of attribution towards AI, the approach outlined in this paper has discovered novel agents and factors people are considering when faced with the AI "responsibility gap". The method can be further used for exploring new and replicating old questions pertaining to people's blame attributions (Bender, 2020). This can be done with readily available, large-scale datasets of tweets. The approach proposed in this paper can also serve as a way of avoiding entrenched measurement bias within a specific field (Oort, Visser, & Sprangers, 2009).

# References

Bender, A. (2020). What is causal cognition? *Frontiers in psychology*, *11*, 3.

Franklin, M., Ashton, H., Awad, E., & Lagnado, D. (2022). Causal framework of artificial autonomous agent responsibility. In *Proceedings of the 2022 aaai/acm conference on ai, ethics, and society* (pp. 276–284).

Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *Iscience*, *24*(4), 102252.

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers*, *20*, 531–558.

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2009). Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of clinical epidemiology*, *62*(11), 1126–1137.

Ruhl, K. (2004). Qualitative research practice: a guide for social science students and researchers. *Historical Social Research*, *29*(4), 171-177. doi: https://doi.org/10.12759/hsr.29.2004.4.171-177

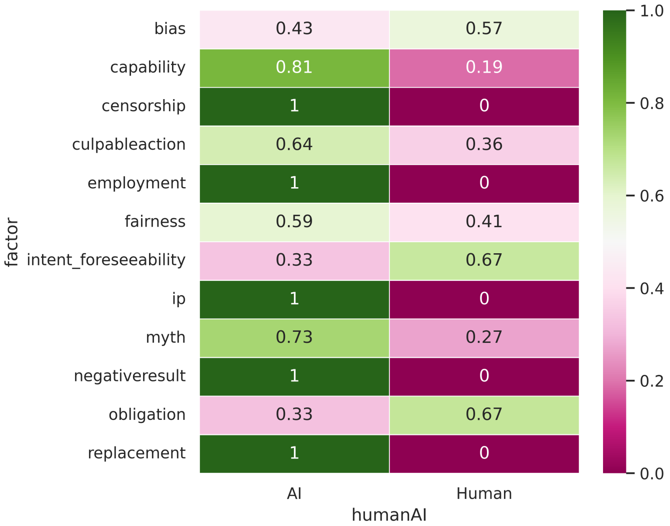

Figure 1: Prevalance of factors in human-AI and human-only contexts