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

Matija Franklin, Trisevgeni Papakonstantinou, Tianshu Chen, Carlos Fernandez-Basso, David Lagnado
University College London, Causal Cognition Laboratory,
London, UK

Abstract

We introduce a novel methodology to scrutinize blame attributions in 'Tweets', focusing on Artificial Intelligence (AI) incidents - a contemporary issue that provokes regular discourse. The method identifies the agents that get blamed and the factors that are associated with blame attributions. The proposed methodology replicates and contextualizes findings from experimental settings, revealing AI entities are often held accountable for adverse outcomes, while human agents are judged based on intentions. It also identifies unexplored factors, such as blaming data for perceived biases or AI for replacing humans. This method offers a robust tool for mitigating measurement bias in specific fields, enabling the continual rejuvenation of theoretical frameworks with emerging variables.

Keywords: Blame; Attribution; Artificial Intelligence; Twitter

Introduction

This paper proposes and tests a method to study people's blame attributions publicly stated on online platforms. It focuses on the blame attributions displayed in "tweets" of Twitter users, reacting to different incidents caused by Artificial Intelligence (AI). Twitter was chosen as it offers an *Academic Research Product Track*, which provides researchers with free historical data of discourse which took place on the platform. AI Incidents were chosen as they are a contemporary topic that is often talked about on Twitter, and currently, the focus of many academic papers due to the nature of *The Responsibility Gap* present when an AI does something that might be blameworthy (Santoni de Sio & Mecacci, 2021). These issues have recently been investigated empirically using online experiments (Rahwan et al., 2022). This present paper proposes a more *ecologically valid*[1] method that may replicate findings from this area of research. The benefit of the approach is the potential discovery of novel factors that haven't been traditionally manipulated or measured in experimental settings. The outlined method can also be applied to different online platforms, and research topics within the field of attribution (Bender, 2020).

Blaming Artificial Intelligence

Although an AI may have causal efficacy, it is not clear who should be held responsible when it makes a mistake (Johnson & Verdicchio, 2019). Developers and users cannot fully predict an AI's behaviour and it is not always clear why an AI made a certain decision (Matthias, 2004). In this sense, the responsibility gap is an example of the *problem of many hands*, where multiple different agents bring about an outcome, making each agent's responsibility less apparent (Slota et al., 2021). AI as an autonomous agent presents a novel challenge due to the possibility of different principal-agent relationships[2] (Kim, 2020; Ho, Slivkins, & Vaughan, 2016). Specifically, an AI may perform different tasks in various ways. It may also have various ways of responding to human input. Further, the human-in-the-loop can have different levels of oversight over the AI agent. General purpose AI systems (such as GPT-4) complicate this further as agents "that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained (Gutierrez, Aguirre, Uuk, Boine, & Franklin, 2022)."

Prior research on how people attribute responsibility[3] has identified factors that influence people's judgments across contexts (D. Lagnado & Gerstenberg, 2015; Vincent, 2011; Perry, 2000). Franklin, Ashton, Awad, and Lagnado have proposed a framework outlining nine factors that have causal influence over responsibility attributions - *causality*, *role*, *knowledge*, *objective foreseeability*, *capability*, *intent*, *desire*, *autonomy*, and *character*.

While an agent can cause an outcome but not be blamed for it, causality is a precursor to attributing responsibility (D. Lagnado & Gerstenberg, 2015). Agents are responsible for carrying out actions according to their role (Gibson & Schroeder, 2003) and are blamed more highly for highly foreseeable outcomes, which relates to their knowledge (D. A. Lagnado & Channon, 2008). Objective foreseeability, which represents how likely an outcome is irrespective of what an agent subjectively foresees, also affects blame attributions. Expectations of an agent's capability influence blame attributions. High expectations of capability result in more blame for negative outcomes (Gerstenberg et al., 2018). Furthermore, intentionality influences blame attributions because they allow one to identify the effects an agent intended (Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015). Desire is conceptually different from desire in that intention involves committing to performing an intended action (Malle, 2001) and also influences blame attribution (Cushman, 2008). Finally, people blame more autonomous agents as they have more control over their own decisions (Alicke, 2000).

People are willing to attribute responsibility to AI (Franklin, Awad, Ashton, & Lagnado, 2023). Research has

[1] Ecological validity in psychological research refers to the extent to which the findings of a study can be generalized to, and are representative of, real-world conditions.

[2] Principal-agent relationships refer to a contractual arrangement where one party (the principal) legally delegates authority to another party (the agent) to act and make decisions on its behalf.

[3] Resposibility relates to *outcome responsibility* - people's attributions of blame or praise for actions that have occurred in the past

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

identified patterns in how people judge AI. People judge AIs more for the outcomes of their actions, and humans more for their intentions (Hidalgo, Orghian, Canals, De Almeida, & Martin, 2021). AIs are blamed more for causing physical harm, and humans are blamed more when they treat someone unfairly. Discrimination by an algorithm causes less moral outrage than discrimination by a human (Bigman, Wilson, Arnestad, Waytz, & Gray, 2022). People are also more likely to centralize responsibility to a higher authority when an AI makes a mistake (Hidalgo et al., 2021). When people judge human-AI teams, they attribute less blame and causality to the AI when both agents make an error (Awad et al., 2018). Further, people receiving advice from an AI get more blame than people receiving advice from a human (Westcott & Lagnado, 2019).

People's perceptions of AIs' capability influence the way they judge and interact with them. People expect people to make mistakes and automation to be flawless (Madhavan & Wiegmann, 2007). In turn, people are less willing to excuse machines for mistakes (Hidalgo et al., 2021). People will also rely more on algorithmic advice as task difficulty goes up (Bogert, Schecter, & Watson, 2021). People are less trustworthy of AI when dealing with tasks that are subjective (Castelo, Bos, & Lehmann, 2019) or anything that involves emotions (Waytz & Norton, 2014). Finally, people prefer not to use artificial autonomous agents for making moral decisions (Dietvorst & Bartels, 2021). They expect artificial autonomous agents, to make utilitarian moral choices, and blame them when they don't (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015). These findings may be explained by the fact that people perceive machines as agents that cannot fully think or feel (Bigman & Gray, 2018; Ashton & Franklin, 2022b), or as agents that are selfish and uncooperative (Ishowo-Oloko et al., 2019).

People's perceptions of AIs' autonomy also influence their judgments. AIs can be viewed as more or less autonomous, with people inferring more intent towards more autonomous AIs (Banks, 2019). Further, robots described as autonomous received blame attributions that were nearly equal to attributions towards humans (Furlough, Stokes, & Gillan, 2021). Similarly, people using autonomous technologies received less praise as they were seen as having less control over these technologies (Jörling, Böhm, & Paluch, 2019). Finally, drivers of automated vehicles are seen as less responsible than drivers of manual vehicles (McManus & Rutchick, 2019).

An issue in cognitive and psychological research is that researchers can only model the factors that get measured. Deciding what gets measured is heavily influenced by the research history of the field (Kapoor et al., 2018). A noteworthy example of this comes from the relationship between people's perceptions of intent, capacity, and blame. Perceptions of the relationship between these factors vary greatly between different academic traditions. The relationship between intent and skill are not features of any known legal concept (Cushman, 2008). In psychological research, on the other hand, actions are seen as more intentful, and thus more blameworthy, if the agent receiving the blame has the necessary skill to execute that action (Malle & Knobe, 1997). The aim of this research is to provide context for previously researched factors, as well as to identify new factors.

The Present Study

Previous research studying attribution towards autonomous artificial agents has mostly used vignettes (Franklin, Awad, & Lagnado, 2021) or evidence in the form of images (Ashton, Franklin, & Lagnado, 2022). In such studies, certain aspects of these vignettes or images that pertain to factors that influence blame are manipulated. The present paper uses 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.

First tweets are identified as either being an attribution or not. The *agents* people are attributing blame to, and the *factors* that have an effect on people's attributions are then identified. Agents are often context-specific, thus identifying them requires a bottom-up approach - identifying which agents are blamed when an AI makes a mistake. The investigated factors that are highly correlated with blame build on the framework proposed by Franklin et al., examining how these factors are used by people making attributions outside the context of an experimental study.

Taking a computational social science approach can provide new data that contextualizes experimental findings. Social media data comprise of digital traces of human interactions that allow us to unobtrusively observe people's real-life behaviour (Pfeffer et al., 2023). The predominantly written format of online data, which features real-world examples of human behaviour along with personal and network information, enables the application of innovative natural language processing technologies to derive insights into human psychology from language. The method is able to capture naturally-occurring behaviour, rather than behaviour displayed in an experimental setting. It is open-ended, that is, "participants" are not assigned a task - they simply behave as they would. This avoids *measurement bias*[4]. Moreover, it has the ability to capture live reactions to critical events (Kapoor et al., 2018).

Methods

Data collection

We focused our sample on tweets responding to famous incidents involving AI. 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., an algorithm designed

[4] Measurement bias refers to systematic errors in data collection that skew results, often stemming from flawed testing instruments, observer subjectivity, or the influence of the experimenter's expectations on the outcomes.

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). The rationale for selecting these specific contexts is primarily based on the media attention they attracted and thus; thus, having a large number of tweets discussing them. Secondly, they represent a combination of factors already known to influence blame attributions, such as physical harm, fairness-related outcomes, intent, and role-related obligations.

The tweets were extracted using Twitter's official API with an academic license. The tweets were identified using search queries including related keywords. The keywords were selected to be broad to avoid ”fishing” for significant effects with regards to agents and factors; we used keywords referring to the context of the incident and their variations (e.g. ”harvard admission lawsuit”). To reduce the probability of collecting unrelated tweets, the search was limited to one month after the incident. A final sample of 23789 tweets was collected through the API. From that dataset, we took a random stratified sample of tweets for which the analysis is presented in this paper.

All tweets and meta-data gathered, as well as the scripts and keywords used for data collection can be found at `https://osf.io/t6sw3/`.

Analytic strategy

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*. 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.

Qualitative coding We applied a variation of the framework method (Ruhl, 2004), a comparative form of thematic analysis that follows a structure of inductively and deductively-created themes. Five trained coders independently coded a sample of 1342 tweets initially according to whether they involved a responsibility attribution for the pre-specified contexts. The subset of tweets that involved an attribution (N=563) was then coded in terms of the agents the attribution was directed at (e.g., the self-driving car), the factors that were involved in making this attribution (e.g., capability), and the sentiment of the attribution, which could have been either positive, negative, or neutral. When not possible, a tweet's agents, factors, and sentiment would get labeled with none. We started with a pre-specified set of codes, based on well-established findings (see: `https://osf.io/t6sw3/`) and through data familiarisation and calibration, we expanded that initial set. Finally, we grouped the codes into themes. 23% (N=308) of tweets were blindly double-coded and coders followed a cyclical analysis approach culminating in triangulation at the final stage. The inter-rater reliability for the coding of blame attributions was substantial (Kappa = 0.70, p <.001).

Table 1: Agents and factors derived through qualitative analysis

Code	N (%)	Description
Agents		
Algorithm	180 (32%)	AI
Company	75 (13%)	Name of a company or representative of a company
Data	6 (1%)	Data the AI is trained on
Developer	44 (8%)	Developer of the AI
Government	79 (14%)	Government as a whole and specific member
Media	17 (3%)	Media source discussing the incident
Person	52 (9%)	A third party discussing the incident
System	72 (13%)	The system around the main agents, enablers, and barriers
User	36 (6%)	User of AI or equivalent system
Victim	6 (1%)	Victim of an incident
Factors		
Bias	118 (21%)	Expression of prejudice for or against an agent or group
Capability	110 (19%)	Capability to fulfill a role, referencing skill or knowledge
Censorship	25 (4%)	Suppression of speech or information
Culpable action	31 (5%)	An act that is in itself blameworthy, regardless of outcome
Employment	87 (15%)	Use of algorithm
Fairness	34 (6%)	Fairness/unfairness explicitly pointed out using relevant language
Intent/Foreseeability	5 (1%)	Harm that was intended or foreseen
Intellectual property	9 (4%)	Discussion of ownership and intellectual property
Myth	17 (3%)	Misconception about how a system works (e.g. black-boxing)
Negative result	77 (14%)	An unexpected or negative outcome
Obligation	39 (7%)	A moral obligation or duty, generally attached to a role
Replacement	15 (3%)	Discussion of humans being replaced by AI

Statistical approach 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 opted to use k-means clustering as the primary aim of this study was not to simply predict blame but rather to explore how factors, agents, and sentiment group together to form an attribution, in a bottom-up manner.

Results

Qualitative

We identified 10 agents and 12 factors related to attribution across the different contexts. Table 1 presents the codes and their descriptions, along with their prevalence in the dataset.

Agents with a high proportion of blame attributions were the algorithm (28%), company (13%), government (12%), and system (11%). Factors with a high proportion of blame attributions were bias (20%), negative result (16%), and capability (15%). Figure 1 presents the allocation of blame across factors and agents.

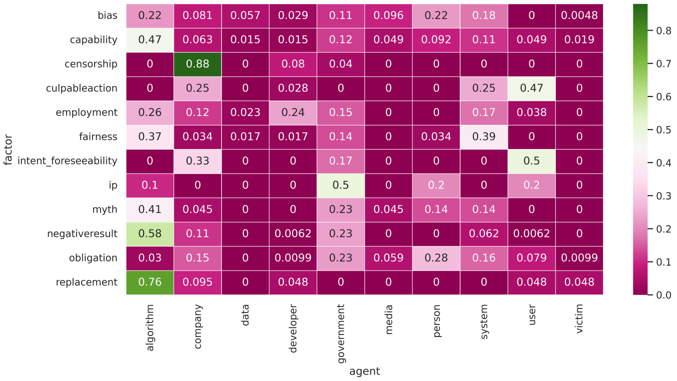

Figure 1: Percentage of blame attributed to pairs of agents and factors

Figure 2 presents the allocation of blame across agents and factors. 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%).

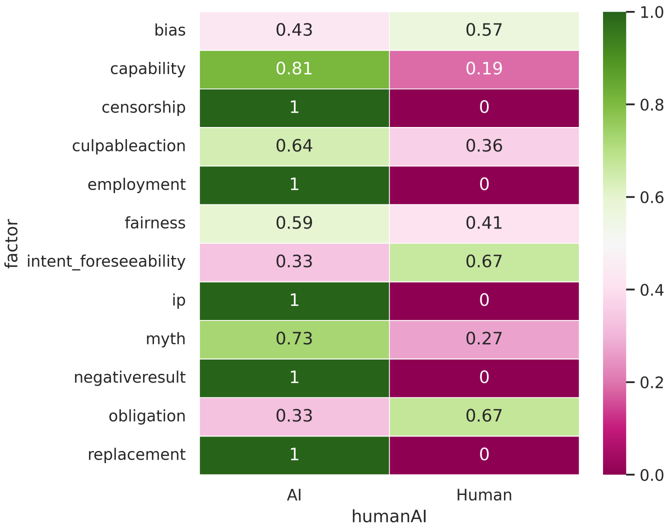

Figure 2: Prevalance of factors in human-AI and human-only contexts

Quantitative

Associations Figure 3 presents the correlation coefficients for all pairwise combinations of blame, context, sentiment, agents, and factors. Blame was most strongly positively correlated with human-only scenarios ($r(703) = .22$, $p = <.001$), negative sentiment ($r(703) = -.72$, $p = <.001$), and bias ($r(703) = .21$, $p = <.001$). It was strongly negatively correlated with neutral ($r(703) = -.53$, $p = <.001$) and positive ($r(703) = -.49$, $p = <.001$) sentiment, employment ($r(703) = -.28$, $p = <.001$), and replacement ($r(703) = -.20$, $p = <.001$).

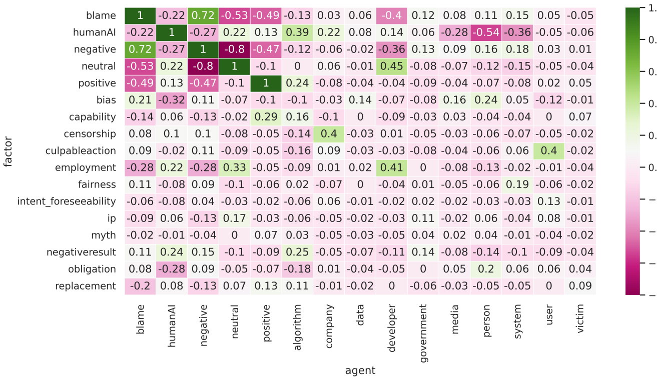

Figure 3: Pearson’s r correlation coefficients for all variables of interest

Cluster analysis The k-means cluster analysis grouped the variables into 6 clusters. Clusters 3, 5, and 6 grouped together tweets containing >94% blame attributions. The sentiment in those clusters was almost entirely negative (>93%). The agents blamed in those scenarios ranged across the codes, with no particular agent being prevalent. Cluster 5 grouped together tweets with the negative result as the main factor relevant to the attribution. Cluster 6 grouped together tweets with bias as the main factor. Clusters 2 and 4 both had a low percentage of blame attributions (<15%). In Cluster 2 the main agent receiving the attribution was the algorithm, and the sentiment was split between neutral and positive. Cluster 4 represented a small percentage of observations and grouped together tweets with neutral sentiment and the main factor of employment. Finally, Cluster 1 grouped together all tweets with praise attributions, as well as many with blame attributions (79%). The sentiment was mostly negative and the main factor relevant to the attribution was capability. Table 3 presents the qualitative features of the clusters in detail. Figure 4 presents the t-SNE projection of the clustering.

Discussion

Twitter users make attributions towards agents and enrich these attributions with relevant factors. Such tweets are a valuable source of data for identifying which agents the attributions are directed towards, as well as the relevant factors contained within them. This expands on research that has used Twitter to research sentiment and opinion (Pak, Paroubek, et al., 2010; Giachanou & Crestani, 2016; Khan, Bashir, & Qamar, 2014), and provides a new tool for researching attribution (Bender, 2020; Franklin et al., 2022).

Arguably, social media users were able to grasp the complexity of “the responsibility gap” (Santoni de Sio & Mecacci,

Table 2: Cluster descriptions

	Cluster 1	Cluster 2
Prevalence	21%	20%
Attribution	79% blame, 16% praise	12% blame
Sentiment	79% negative	61% neutral, 19% positive
Agents	69% algorithm	52% algorithm, 12% government, 12% user
Factors	72% capability, 13% employment	22% bias, 20% replacement, 15% IP
	Cluster 3	**Cluster 4**
Prevalence	27%	6%
Attribution	96% blame	15% blame
Sentiment	94% negative	100% neutral
Agents	43% company, 19% system, 13% government	71% developer, 17% company
Factors	split across	96% employment
	Cluster 5	**Cluster 6**
Prevalence	13%	13%
Attribution	94% blame	100% blame
Sentiment	93% negative	100% negative
Agents	69% algorithm, 25% goverment	27% person, 26% algorithm, 18% system
Factors	100% negative result	100% bias

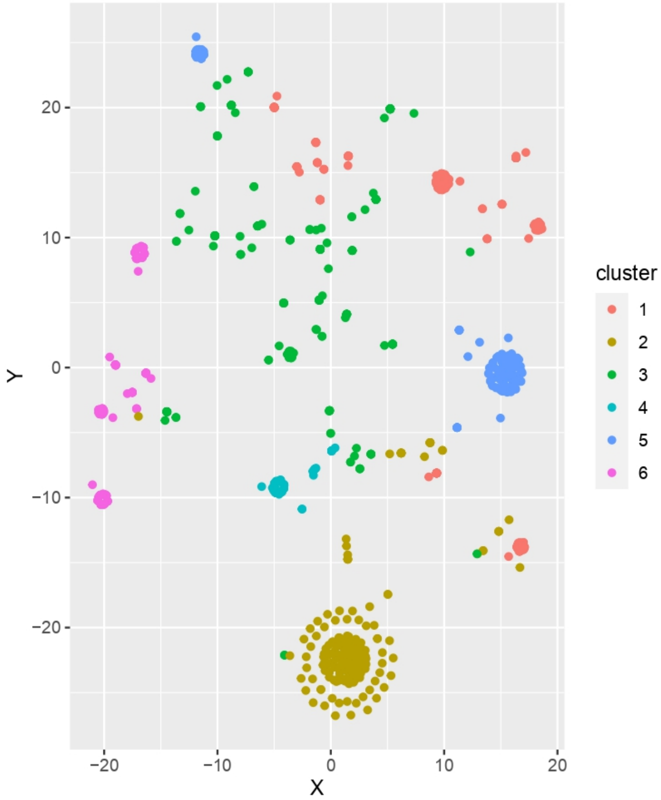

Figure 4: t-SNE projection of dataset (all corpora)

2021), making attributions toward 10 different agents that brought about the outcome, and 12 factors that were relevant to them. The most blamed agents were algorithm (28%), company (13%), government (12%), and system (11%), respectively. This replicates the finding that people are willing to make attributions towards algorithms directly, as well as to other agents within a context (Franklin et al., 2022). As with past research participants also tend to centralize responsibility to the nearest human with authority when an AI makes a mistake (Hidalgo et al., 2021). The government also got blamed which is an agent that hasn't been considered in experiments on AI blame. The way people blame *group agents*, such as governments or companies, has been explored in the literature (List & Pettit, 2011) and discussed in relation to AI (List, 2021). People are willing to view group entities as agents, and this shapes the way they think about and interact with them.

People were also willing to blame the broader system around the main agents. Attribution toward systems is an understudied phenomenon in previous research. People are indeed capable of thinking about systems (Meadows, 2008), but when, why, and how they blame them remains an open question.

The most commonly used factors to blame were bias (20%), negative result (16%), and capability (15%), respectively. Although bias in AI has been extensively researched by AI Ethics researchers (Ntoutsi et al., 2020; Ferrer, van Nuenen, Such, Coté, & Criado, 2021; Yapo & Weiss, 2018), its impact on attribution has been seldom explored. The finding that people focus on negative outcomes when making attributions replicates previous research (Hidalgo et al., 2021). This is also true for capability (Gerstenberg, Ejova, & Lagnado, 2011).

The percentage of blame attributed to pairs of agents and factors in Figure 1 reveals certain patterns. Algorithms are often blamed for causing negative outcomes, in line with research showing that machines are blamed more than humans for the outcomes of their actions (Hidalgo et al., 2021). They are also often blamed for replacing humans, which similarly to bias has often been a topic of research in AI Ethics (Ashton & Franklin, 2022a; Vorobeva et al., 2022), but less so in attribution research. The companies that make AI were often blamed for using AI to censor certain users, which is adjacent to experimental research on users' judgments of moderation on social media platforms (Myers West, 2018). When data was blamed, it was most likely blamed for the bias contained within it; a topic very often discussed in machine learning research (DeBrusk, 2018). Although anecdotal evidence is available for people's willingness to blame data as an artificial agent, to the authors' best knowledge this is the first research paper to document this phenomenon.

When governments were blamed, they were most often blamed due to broader concerns about ownership and IP. Such attributions were often related to the government's obligation to regulate. This is in line with previous research showing that people are willing to attribute responsibility to a government for outcomes it did not directly cause (Mortensen, 2013). The media was most often blamed for being seen as biased, replicating previous research (Glynn & Huge, 2014). Further, ”the system” was blamed for being unfair. Finally, users of AI were most often blamed for their intent or foresight, thus replicating previous findings (Hidalgo et al., 2021). Users also often got blamed for performing *culpable actions* - acts that are in themselves blameworthy, irrespective of the outcome (e.g., lying or cheating).

The prevalence of the 12 factors in human-AI and human-only contexts is available in Figure 2. Bias, obligation, and intent/foreseeability were more prevalent for human-only systems, whilst capability, negative outcomes, and replacement were more prevalent for human-AI systems. 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). As bias as such is not explored by previous research, it may be the case that Twitter users view bias in humans as a culpable action, whilst bias in machines is more statistical in nature. 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). Finally, it may be the case that replacement is more prevalent in human-AI contexts as there is a current active debate about the extent to which AIs will be able to perform certain tasks better than humans (Brynjolfsson & McAfee, 2014).

Blame was positively associated with negative sentiment and negatively associated with neutral and positive sentiment (see Figure 3). The stronger positive than negative correlation relates to previous findings showing that blame is more differentiated and more extreme than praise (Guglielmo & Malle, 2019). The rest of the associations mirror the patterns previously discussed in Figure 1. Overall, blame exhibited strong (i.e., $>.02$) positive associations with bias and strong negative associations with employment and replacement.

The six clusters also reveal certain unique patterns. Cluster 1 uniquely contains praise attributions, mostly towards algorithms, and mostly for their capability. This was often the case for tweets making attributions towards AI-generated text and art. Clusters 3, 5, and 6 are all large blame clusters (i.e., $>94\%$) and are mostly negative in sentiment. Cluster 5 exclusively contains attributions focused on negative outcomes, while cluster 6 only contains attributions focused on bias. Cluster 3 on the other hand contains a range of different factors. Clusters 2 and 4 are low in blame, which is evidenced by their relatively high neutral and or positive sentiment.

Limitations

This approach outlined in this paper is not without limitations. Naturally-occurring datasets limit the researcher's control and thus the ability to make inferences. Furthermore, the short length of the text that Twitter allowed its users at the time significantly reduces the amount of information that can be directly conveyed. As a result, attributions are not always made in an explicit way. This significantly limits a human coder's ability to discern whether such a judgment is indeed being made, which is even more challenging for a supervised algorithm to do. Exploring new Tweets with longer word counts on Twitter can overcome this limitation.

The main theoretical limitation of this study is the lack of control over the potentially confounding effect of context. It is possible that the different events are qualitatively unique and that the effects of agents and factors correspond to the specific qualitative features of the context, rather than describing mechanisms of attribution more broadly. The association and cluster analyses fail to control for this effect, which potentially undermines the generalisability of our findings to other AI-related incidents. This limitation is due to the structure and complexity of the data – namely the agent and factor variables, which comprise of numerous levels – resulting in models with high degrees of freedom that require a larger volume of observations. Nevertheless, the bottom-up, context-dependent approach that results in this theoretical limitation represents a strength of the study for different reasons, as discussed previously.

Future research can be directed towards applying and examining this framework using richer datasets that consist of a longer text. Data from similar media such as Reddit or other forum-like websites can be used to further interrogate this model. Additionally, the current dataset can be explored in alternative ways that allow for natural themes to emerge from the text without human direction, such as topic modeling. Specifically, the theoretical concern of context-dependence could be addressed by structural approaches to topic modeling that can include such control variables.

Conclusion

The approach outlined in this paper has discovered novel agents and factors people are considering when faced with the AI "responsibility gap". It has also replicated findings from experimental settings, giving them context and strengthening their ecological validity. 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 large-scale nature of this type of research would give researchers more ecological validity to their proposed models. One may find that a factor that was highly significant in the lab, is barely mentioned in certain contexts of public online discourse.

The approach proposed in this paper can serve as a way of avoiding entrenched measurement bias within a specific field (Oort, Visser, & Sprangers, 2009). It does so by allowing one's framework to update itself if this is what is reflected in the analyzed discourse. This study is also a step toward training an algorithm capable of identifying and analysing attributions in social media posts.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556.

Ashton, H., & Franklin, M. (2022a). The corrupting influence of ai as a boss or counterparty.

Ashton, H., & Franklin, M. (2022b). A method to check that participants really are imagining artificial minds when ascribing mental states. In *Hci international 2022–late breaking posters: 24th international conference on human-computer interaction, hcii 2022, virtual event, june 26–july 1, 2022, proceedings, part ii* (pp. 470–474).

Ashton, H., Franklin, M., & Lagnado, D. (2022). Testing a definition of intent for ai in a legal setting. *Submitted manuscript*.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., … Rahwan, I. (2018). Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. *arXiv preprint arXiv:1803.07170*.

Banks, J. (2019). A perceived moral agency scale: development and validation of a metric for humans and social machines. *Computers in Human Behavior*, *90*, 363–371.

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

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*.

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, *11*(1), 1–9.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

DeBrusk, C. (2018). The risk of machine-learning bias (and how to prevent it). *MIT Sloan Management Review*.

Dietvorst, B. J., & Bartels, D. M. (2021). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology*.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80.

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., Ashton, H., & Lagnado, D. (2023). Unpredictable robots elicit responsibility attributions. *Behavioral and Brain Sciences*, *46*, e30.

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

Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. the influence of robot autonomy. *Human Factors*, *63*(4), 592–602.

Gerstenberg, T., Ejova, A., & Lagnado, D. (2011). Blame the skilled. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, *49*(2), 1–41.

Gibson, D. E., & Schroeder, S. J. (2003). Who ought to be blamed? the effect of organizational roles on blame and credit attributions. *International Journal of Conflict Management*.

Glynn, C. J., & Huge, M. E. (2014). How pervasive are perceptions of bias? exploring judgments of media bias in financial news. *International Journal of Public Opinion Research*, *26*(4), 543–553.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one*, *14*(3), e0213544.

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., & Franklin, M. (2022). A proposal for a definition of general purpose artificial intelligence systems. *Available at SSRN 4238951*.

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

Ho, C.-J., Slivkins, A., & Vaughan, J. W. (2016). Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, *55*, 317–359.

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521.

Johnson, D. G., & Verdicchio, M. (2019). Ai, agency and responsibility: the vw fraud case and beyond. *Ai & Society*, *34*(3), 639–647.

Jörling, M., Böhm, R., & Paluch, S. (2019). Service robots: Drivers of perceived responsibility for service outcomes. *Journal of Service Research*, *22*(4), 404–420.

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.

Khan, F. H., Bashir, S., & Qamar, U. (2014). Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision support systems*, *57*, 245–257.

Kim, E.-S. (2020). Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective. *Technology in Society*, *63*, 101378. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0160791X19306906` doi: https://doi.org/10.1016/j.techsoc.2020.101378

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Cogsci.*

Lagnado, D., & Gerstenberg, T. (2015). A difference-making framework for intuitive judgments of responsibility. *Oxford studies in agency and responsibility*, *3*, 213–241.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

List, C. (2021). Group agency and artificial intelligence. *Philosophy & technology*, *34*(4), 1213–1242.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301.

Malle, B. F. (2001). Intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, 45.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology*, *33*(2), 101–121.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *2015 10th acm/ieee international conference on human-robot interaction (hri)* (pp. 117–124).

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175–183.

McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social psychological and personality science*, *10*(3), 345–352.

Meadows, D. H. (2008). *Thinking in systems: A primer*. chelsea green publishing.

Mortensen, P. B. (2013). (de-) centralisation and attribution of blame and credit. *Local Government Studies*, *39*(2), 163–181.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., … others (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

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.

Pak, A., Paroubek, P., et al. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, pp. 1320–1326).

Perry, S. R. (2000). Loss, agency, and responsibility for outcomes: Three conceptions of corrective justice. *Philosophy of law*, *6*, 546–559.

Pfeffer, J., Matter, D., Jaidka, K., Varol, O., Mashhadi, A., Lasser, J., … Morstatter, F. (2023). *Just another day on twitter: A complete 24 hours of twitter data.*

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … others (2022). Machine behaviour. *Machine Learning and the City: Applications in Architecture and Urban Design*, 143–166.

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

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*, 1057–1084.

Slota, S. C., Fleischmann, K. R., Greenberg, S., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2021). Many hands make many fingers to point: challenges in creating accountable ai. *AI & SOCIETY*, 1–13.

Vincent, N. A. (2011). A structured taxonomy of responsibility concepts. In *Moral responsibility* (pp. 15–35). Springer.

Vorobeva, D., El Fassi, Y., Costa Pinto, D., Hildebrand, D., Herter, M. M., & Mattila, A. S. (2022). Thinking skills don't protect service workers from replacement by artificial intelligence. *Journal of Service Research*, *25*(4), 601–613.

Waytz, A., & Norton, M. I. (2014). Botsourcing and outsourcing: Robot, british, chinese, and german workers are for thinking—not feeling—jobs. *Emotion*, *14*(2), 434.

Westcott, C., & Lagnado, D. (2019). The ai will see you now: Judgments of responsibility at the intersection of artificial intelligence and medicine (master's thesis). *Unpublished Manuscript*.

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning.