4-24-2023

Human-AI Collaboration in Content Moderation: The Effects of Information Cues and Time Constraints

Haoyan Li
National University of Singapore, e0998154@u.nus.edu

Michael Chau
The University of Hong Kong, mchau@business.hku.hk

HUMAN-AI COLLABORATION IN CONTENT MODERATION: THE EFFECTS OF INFORMATION CUES AND TIME CONSTRAINTS

Research in Progress

Haoyan Li, National University of Singapore, Singapore, e0998154@u.nus.edu

Michael Chau, The University of Hong Kong, Hong Kong, mchau@business.hku.hk

Abstract

An extremely large amount of user-generated content is produced by users worldwide every day with the rapid development of online social media. Content moderation has emerged to ensure the quality of posts on various social media platforms. This process typically demands collaboration between humans and AI because of the complementarity of the two agents in different facets. Wondering how AI can better assist humans to make final judgment in the "machine-in-the-loop" paradigm, we propose a lab experiment to explore the influence of different types of cues provided by AI through a nudging approach as well as time constraints on human moderators' performance. The proposed study contributes to the literature on the AI-assisted decision-making pattern, and helps social media platforms in creating an effective human-AI collaboration framework for content moderation.

Keywords: Content moderation, Human-AI collaboration, AI-Assisted decision-making, Information cues, Nudging, Debiasing.

1 Introduction

Online social media has achieved rapid development worldwide in the past decade. Large amounts of user-generated content (UGC), including text, images and videos, are produced daily by millions of social media users around the globe. While bringing useful information and entertainment to people, UGC also allows harmful information to be rapidly disseminated, which may lead to devastating consequences such as social instability. As the service providers and the profit gainers, the user-generated content platforms and commercial websites should take the responsibility to provide a safe and trustworthy environment for their users, complying with local regulations, policies and social norms.

Content moderation is an organized approach which screens UGC published on websites, social media, and other online channels to determine whether the content is appropriate for a specific website, region, or jurisdiction. The moderation style can be highly diversified among different sites and platforms because the rules of what content is allowed are usually set at a site or platform level, and reflect the brand and reputation, risk tolerance magnitude, and the ideal type of user participation of this platform (Roberts, 2017). The two objectives of content moderation are (1) fulfilling social responsibility by preventing the spread of inappropriate content, and (2) improving user experience to the largest extent by endowing users with great freedom of speech within the scope of morality and legislation. Therefore, the decision rule should be set at an appropriate level: if the benchmark is too lax, the chance of harming content being released to the public would increases; whereas if the benchmark is too stringent, users' enthusiasm to generate content will be dampened.

Nowadays, most social media companies such as Facebook, Twitter, YouTube, and TikTok adopt the content moderation strategy which allows humans and AI working together to identify, flag, hide, or remove potential toxic content. It is impossible for human moderators to classify millions of posts each

day manually with limited time and budgets. It could also be hard for them to describe their decision rules very accurately (Fügener, Grahl, Gupta & Ketter, 2021), which may result in inconsistency in decision-making among different moderators when handling similar cases. Even for the same moderator, decisions could be inconsistent in different time periods due to external factors such as fatigue, distraction, and emotional fluctuations.

Never interfered by external factors that humans frequently suffer from, machines are more objective, and decisions are always made based on the same rule with a high degree of consistency (Sundar, 2008; Binns et al., 2018). However, machines may be viewed as incompetent in making nuanced subjective judgments with full automation (Molina & Sundar, 2022) such as on ethical (Awad et al. 2018) and legal (Kingston, 2016) challenges.

Considering the respective strengths of humans and AI in different facets, their roles in completing a certain task can be complementary. There are already some studies demonstrating that collaboration between AI and humans outperforms the AI agent or the human agent working alone in certain scenarios (Fügener, Grahl, Gupta & Ketter, 2021). Not as a rare phenomenon, AI-human collaboration has been increasingly applied to many high-stake domains such as medical diagnosis, candidate screening for hiring, and loan approvals. As a formulation of teaming, this type of human-AI collaboration incorporates a binary trust model, in which humans either trust or distrust the output produced by AI. (Bansal et al., 2019). However, the practice of content moderation is different from other typical human-AI-collaborative decision-making tasks for it incorporates an extremely large volume of continuous decisions rather than discrete individual decisions (Lai et al., 2022). In addition, the moderation system is required to complete the massive classification tasks within a very short time duration, typically a few seconds for each case, which is particularly challenging for human moderators. Therefore, the collaboration pattern between the AI agent and the human agent can be critical in determining the accuracy and efficiency of the entire decision system.

Focusing on the context of content moderation of user-generated content on social media, the primary purpose of this research-in-progress paper is to develop an experimental design to investigate an efficacious human-AI collaboration pattern concerning tasks with high massiveness and timeliness. Based on this general objective, our research question is: How can AI assist humans to make better decisions in content moderation? To answer this question, we draw upon the literature on nudging and debiasing and introduce three types of information cues generated by AI as a nudging strategy to assist humans in the "machine-in-the-loop" collaboration process. To be specific, we would make comparison between implementing each cue alone and three cues all together. Furthermore, considering the common pressing time constraints in the commercial content moderation scenario, we are interested in the interaction effect between nudging and time constraints. Thus, our second research question is formulated as following: What is the interaction effect between nudging and time constraints in such an AI-assisted decision-making workflow? Our proposed experimental design is on the basis of these two research questions.

2 Literature Review

2.1 Two Paradigms of Human-AI Collaboration

Several studies have found that humans and AI working together can yield better performance than AI and humans when working alone in specific contexts. Nevertheless, while algorithms can often make predictions more accurately than humans, their inability to ratiocinate and adjust to novel or marginal situations makes them unsuitable for implementing many principles of responsible and ethical decision-making (Alkhatib & Bernstein, 2019). Therefore, the engagement of both humans and AI are indispensable in many types of tasks, and the way in which the two agents collaborate is critical. Here we explore two representative paradigms of human-AI collaboration: "human-in-the-loop" and "machine-in-the-loop", which are of divergent operation mechanisms.

Being widely used, the term "human-in-the-loop" often refers to "the interactive training paradigm where the AI receives input from the human to improve its performance" (Lai et al., 2022). Utilization

of machine-learning technologies in daily life is greatly limited to skilled practitioners for it demands a certain degree of specialized knowledge. Therefore, typical end-users' involvement is largely tuned by those practitioners (Amershi, Cakmak, Knox & Kulesza, 2014). Different from traditional machine learning, the workflow of interactive machine learning is designed to involve user input but does not require experience or the background knowledge in machine learning algorithms (Dudley & Kristensson, 2018). Besides, there are more intent, enhancive and rapid model updates involved in the learning cycles of interactive machine learning processes, allowing users to inspect the effect of their actions and adjust succeeding inputs to acquire desired behaviors in an interactive way (Amershi, Cakmak, Knox & Kulesza, 2014).

In this workflow, annotation, the process of labeling raw data, is critical in providing high-quality training data for AI to refine its classification algorithm. Though knowledge on building machine learning models is not indispensable, annotation work is actually still closely related with the mechanism of machine learning (Monarch, 2021). It is very hard for a large group of moderators to learn annotation skills with the underlying logic in a short period. Another potential problem involved in this workflow is that a user might not strictly stick to a concept during the training process or might bring their own biases in the learning process, which could frequently happen in moderation work involving great subjective judgement and result in errors, whereas correcting those errors requires astonishingly sophisticated statistical techniques (Monarch, 2021). Besides, considering moderators' lack of knowledge in the underlying logic of machine learning, it is a considerable challenge for both the machine learning practitioners and for the user interface designers to construct such a system for moderators which requires transforming the data inspection and correction task into a procreative and intuitive interlocution between humans and machines (Dudley & Kristensson, 2018).

Because of the issues of "human-in-the-loop", another paradigm, "machine-in-the-loop", would be extremely demanding for the massive ordinary moderators to use and update in the near future. "Machine-in-the-loop", also known as "AI-assisted/advised" decision-making, is another human-AI collaboration framework in which machines play a supporting role to achieve the objective of improving the ability of humans. Humans are the central actors and have discretion and responsibility in deciding what to do with machine outputs as the final decision (Clark, Ross, Tan, Ji & Smith, 2018; Zhang, Liao & Bellamy, 2020). Rather than their impact on human decision-making, research and debates concerning algorithmic decision aid made primary emphasis on the statistical properties of models (Angwin, Larson, Mattu & Kirchner, 2016; Dieterich, Mendoza & Brennan, 2016).

While institutions are progressively adopting machine learning models attempting to be "evidence-based" (Sonja, 2014), there is little known about how machine learning models influence decision-making practically. This phenomenon of lacking evidence is particularly troubling given studies showing that people have difficulty interpreting machine learning models and incorporating algorithmic predictions into their decisions, which frequently results in unexpected and unfair outcomes from machine learning systems (Green & Chen, 2019). Fügener, Grahl, Gupta & Ketter (2021) indicate two factors that are related to the ability of humans to obtain benefits from AI advice: relative performance and the ability to distinguish between correct and incorrect advice. Therefore, besides achieving the objective of outstanding performance, security and fairness, machine learning decision-support applications should also be enhanced to help decision-makers comprehend the predicted outputs made by the model.

2.2 Content Moderation

There are only a few existing studies concentrating on the topic of content moderation in the IS-relevant area. One particular perspective is to enhance humans' trust in AI by incorporating more transparency in the moderation task. This is not an easy undertaking due to the challenges of AI handling ethical content as mentioned previously, as well as users' concern that AI takes away the control that they expect (Devito et al., 2017). To solve the problem of distrust, Sundar et al. (2015) suggest two routes in the HAII-TIME model of human–AI interaction: the cue route and the action route (Sundar, 2020). When following the first route, AI could serve as a source cue. Though users may not necessarily make any action, the cue could potentially trigger various heuristics (Sundar, 2008; Wang, 2021). Through a

lab experiment, Molina & Sundar (2022) show that when a system triggers positive heuristics in the users' minds and provides the user with the opportunity to participate in content classification, the users' trust in the AI for content moderation and agreement with the system is enhanced. One approach following this process is to let users suggest the scope of inclusion or exclusion of the words from the classification system. Besides, Suzor et al. (2019) provide further context for calls for greater transparency in content moderation as a communication process to hold independent stakeholders accountable. Another track of studies concerning content moderation is conditional delegation. Lai et al. (2022) propose conditional delegation as an alternative paradigm aside from the human-AI collaboration paradigms "human-in-the-loop" and "machine-in-the-loop" which have been more familiar to people. Conditional delegation combines AI models and traditional rule-based approaches by empowering moderators with the competence to decide when to trust or distrust the AI model. The researchers used a rationale-style neural architecture as a classifier which embraces both approaches to demonstrate that if users are capable of making good selections of keywords rules, conditional delegation would be able to outperform both the model working alone and the manual rule-based approach.

To the best of our knowledge, in the context of content moderation, there is currently no study specifically examining how AI can better assist humans to make final decisions by debiasing through the nudging approach and how time constraints can influence humans' performance in this particular decision-making process. The current study aims to fill this research gap.

3 Hypothesis Development

At present, most social media platforms adopt the content moderation strategy which firstly sends all posts to the AI algorithms and transferring a small portion of them which cannot be precisely classified by AI to human moderators to make the final judgment. This is a typical human-AI task allocation workflow.

As the first content detection step, AI would calculate a probability of risk based on a series of algorithms, and the classification decision given by AI depends on whether the probability exceeds a certain threshold determined by experts. However, there are a number of cases that fall into the "intermediate zone" between the upper threshold of the negative class and the lower threshold of the positive class, representing ambiguous cases which AI does not have enough confidence to classify as either a certainly safe post or a certainly inappropriate post that contains some level of detrimental information. There are several reasons leading to this result. First, the content could be out of the scope of the AI knowledge. Second, it might demand a higher level of subjective fuzziness in the judgment. Third, there may be great variations in the different cases. In this situation, AI would hand over the ambiguous case to human moderators to be further processed based on human knowledge, such as linguistic, cultural, or ethical knowledge. However, AI may continue to engage in the second step by providing assistance to the human moderators. In this case, the workflow is a "machine-in-the-loop" human-AI collaboration process, and the essence of this process in content moderation is to improve AI's capability to better assist humans in the decision-making process.

Because of the time pressure and massive features of tasks in the process of content moderation, very often moderators have to make quick and intuitive judgment based on restricted information. This kind of judgement could be incomplete, biased or ambiguous (Hogarth, 2001). Bias is defined as a deviation from an objective criterion, such as a normative model (Baron, 2012). Debiasing helps people make better use of the information they have by reframing that information in a way that highlights its importance or corrects a misunderstanding (Soll, Milkman & Payne, 2015). There are two typical ways to do debiasing: influencing the decision makers, which refers to assisting people to overcome their restrictions and tendencies by offering them tools, training and knowledge (Arkes, 1991; Larrick, 2004); and modifying the environment, which aims to change the environment to provide a better match for people to think naturally without help (Klayman & Brown, 1993) or as an alternative, to inspire better ideas (Soll, Milkman & Payne, 2015). In the scenario of modifying the environment, a nudge is an intervention that achieves modification that neither limits choice nor alters stimulus significantly (Thaler & Sunstein, 2008), but affects behavior for the benefit of the individual or the society relying on

psychological principles (Soll, Milkman & Payne, 2015). To be specific, if a person anticipates a common underlying error, she or he can serve as "a choice architect" and constitute an environment to "nudge" choices in wise directions through the active design of the decision-making environment (Thaler & Sunstein, 2008). The way of presenting alternatives or information to decision makers in this process is defined as choice architecture (Johnson et al., 2012; Thaler & Sunstein, 2008).

All technology interfaces can be conceived as choice architecture. Human behaviors can be influenced by presentation, consideration, or arrangement of interface features or cues (Lee, Kiesler, & Forlizzi, 2011; Salganik, Dodds, & Watts, 2006), which, therefore, function as nudges (Hou, 2017). If people have appropriate cues as the right information packaged in an intuitively intelligible and persuasive format, there would be a higher chance of them drawing accurate conclusions (Soll, Milkman & Payne, 2015).

In the human-AI-collaboration content moderation context, there are three potential information cues that can be provided by the AI agent to assist human moderators to improve their decisions: (1) the predicted probability of a certain post being classified as an inappropriate post, (2) highlighting of specific keywords in a post that are potentially or ambiguously inappropriate, (3) the predicted subjects as labels of a post (such as "hate speech" or "racism"). For ease of exposition, we use "probability cues", "highlighting cues", and "label cues" to denote the three information cues respectively. We hypothesize that these three cues can improve humans' performance:

H1: Information cues provided by AI can enhance humans' decision-making performance in content moderation through a nudging approach.

As mentioned above, the output of AI in the first moderation step is the probability of risk or a relevant evaluation score. It is the most intuitional type of information cue for human moderators to make decisions in the next step. However, according to Hoffrage, Lindsey, Hertwig, & Gigerenzer (2000), probabilistic information is arrantly puzzling. Alternative representations would help decision makers better comprehend the underlying structure of the problem in an ideal way (Barbey & Sloman, 2007). Particularly, resorting to visual displays is a promising method to deliver information (Galesic, Garcia-Retamero & Gigerenzer, 2009). Moreover, according to Thaler & Sunstein (2008), a key for nudging to be effective is that heuristic cues must be salient. Salient cues are prominent, novel, or easily thought-of information in the environment (Fiske & Taylor, 1991). Besides being drawn forth by novelty and unexpectedness, salience can also be created by drawing people's attention to the cues (Dolan, Hallsworth, Halpern, King & Vlaev, 2010). Based on this idea, we propose highlighting the words that are suspicious of expressing inappropriate information as our second type of cues, a salient visual aid. Besides, the AI agent can also generate some predicted labels indicating the subject of a post according to the topic or the salient elements involved. Some typical labels include "racism", "hate speech", "suicidal ideas", "eroticism", etc. This type of cue provides an intuitive hint to people about the category of harmful information potentially underneath the content. Therefore, we hypothesize the following:

H2: Compared with probability cues, (a) highlighting cues and (b) label cues can better improve humans' decision-making performance in content moderation.

When many people provide judgments, weighing those judgments on an equal scale is a highly efficient method to combine them. Examples include simple averages and majority rule (Clemen, 1989; Hastie & Kameda, 2005). Researchers found that using combinations of classifiers could remarkably improve the classification performance. As the simplest combination method among all, majority vote doesn't require any prior knowledge of the behavior of individual classifiers or a large number of trainings of classification results (Lam & Suen, 1997), which is easy to implement. People have applied the idea of utilizing the "wisdom of crowds" widely in various contexts, such as sports prediction markets and national security (Surowiecki, 2004). Inspired by this idea, we propose a further improved method to

facilitate decision-making, which is aggregating the decisions made with the assistance of the three types of cues by majority vote.

H3: Aggregating all of the three decisions based on the three types of cues can better enhance humans' decision-making performance than any type of cue alone in content moderation.

Due to the large volume of UGC, human moderators are often required to process a great number of cases with very restricted time limits. For this consideration, we also desire to tackle the influence of time constraints on human moderators in this human-AI collaboration context through an interaction effect with nudging.

Many studies demonstrate that time constraints have a negative effect on decision-making. Faced with time-constrained decisions, people have the tendency to do less analytical processing, take fewer options/attributes into consideration, and ascribe more weight to negative information, all of which lead to worse outcomes (Ordóñez, Benson & Pittarello, 2015). As a matter of fact, Inbar, Botti, and Hanko (2011) show that people believe the lay theory "a quick choice is a bad choice". They found that individuals feel rushed under the circumstance that they have to choose from a large number of options, which results in decision regret. Nevertheless, some evidence also indicates that time constraints can positively affect judgments and decision-making by reducing decision biases. It is suggested by the findings on "adaptive decision maker" that people are flexible enough to handle time constraints by minimizing effort without seriously downgrading the quality of decisions (Ordóñez, Benson & Pittarello, 2015). Svenson and Benson (1993) attest that under time constraints, framing bias is weak. Dhar, Nowlis, and Sherman (2000) ratiocinate that context effects arising from excessively focusing on the relational characteristics of the alternative options would be decreased by time constraints. Given the two opposing potential interaction effects of time constraints and nudging on decision-making supported by theoretical arguments, we develop the competing hypotheses:

H4a: The interaction between nudging and time constraints have a positive effect on humans' decision-making performance in content moderation.
H4b: The interaction between nudging and time constraints have a negative effect on humans' decision-making performance in content moderation.

4 Experiment Design

4.1 Procedure and Variables

Following the nudging perspective, this study investigates how interface features generated by AI would affect human moderators' decisions on whether to classify a post as inappropriate. We propose a laboratory experiment with a 4 (no cues, probability cues, highlighting cues, and label cues) × 2 (time constraint: with versus without) between-subject design to test the influence of various information cues provided by AI and time constraints on humans' performance. The materials we will use are user-generated text posts from social media platforms. We will invite 2 to 3 senior experts on content moderation to review the posts, guaranteeing each piece of them is somewhat ambiguous and then make careful classification decisions as our judgment standard.

All of the participants will be informed that during an experiment concerning content moderation, they need to view 50 social media posts, make judgments on whether each post contains inappropriate information or not, and try to finish all the tasks as soon as possible. A copy of the tutorial will be distributed to them, providing detailed instructions on the experiment. Besides, the participants will also be told that an extra reward shall be offered to them based on their performance, so they would have the motivation to complete the experiment to their best.

Before the major experiment, we will use a few attention and knowledge checks to screen out the participants who do not have qualified knowledge in sensitive topics, understanding of social media, or

English language proficiency. Subjects who have scores beyond a set of thresholds will formally participate in the major experiment. Demographic information regarding age, gender, race and education will also be collected at this stage.

Qualified participants will be assigned randomly to the eight conditions of approximately the same size. We will conduct a balance check to observe whether there are significant differences in the values of demographic information and scores of pre-experiment checks across different groups. All participants across the eight conditions will be shown the same 50 UGC posts that contain a certain level of ambiguity, meaning that classification decisions are kind of subjective instead of cut and dried. The posts will cover a range of various topics, including but not limited to racism, hate speech, suicidal ideations, violence, eroticism and terrorism. Each post contains approximately 80 to 150 English words.

An experimental content moderation system based on a sequence-to-sequence with attention model will be designed to simulate the AI agent for processing the posts and generating the probability cues, highlighting cues, and label cues. Widely applied to automatic speech recognition, sequence-to-sequence with attention models are typically trained to optimize the cross-entropy loss function, corresponding to enhancing log-likelihood of the training data (Prabhavalkar et al., 2017).

Furthermore, in the treatment groups with time constraints, the participants will be informed that there is a time limitation of 30 minutes. The time countdown will also be displayed on the experiment interface as a reminder. Anyone who completed the experiment exceeding this time limitation will lose the opportunity to receive the extra reward.

Dependent Variables: There are two metrics to evaluate the performance of human participants: accuracy and efficiency. Since in practice, the misclassification of false negative has a more devastating influence than a false positive case, we will assign a higher weight to false negative classification in the measurement of accuracy. Therefore, accuracy is defined as $1 - (\text{false positives} + \text{false negatives} \times 2) \div 150$. Efficiency will be measured as the length of time (in seconds) a participant spends to finish all of the tasks. The shorter the time length, the higher the efficiency.

Control Variables: This study will control for English proficiency, attention, as well as knowledge of social media and various sensitive topics. We plan to measure English proficiency by *Fill-in-the-Blank Questions (FBQs)* which are widely used from the classroom level to far larger scales to measure participants' proficiency in English as a second language (Sumita, Sugaya & Yamamoto, 2005). We will adopt the three attention checks designed by Lai et al. (2022) to measure participants' ability to comprehend the instructions and focus on the experiment. They will be given trial questions with a few simple posts to label if they are appropriate or not based on our definition. Besides, we will design a test which contains multiple-choice questions on the features and functions of major social media, social news concerning sensitive topics and buzzword explanations to measure participants' knowledge of social media and sensitive topics.

4.2 Planned Data Analysis and Expected Results

The measurement items are accuracy and efficiency. We will conduct a two-way MANCOVA with accuracy and completion time as dependent variables to examine the overall difference among groups with different types of information cues and with or without time constraints. We expect the scores for age, gender, education, race, attention, English proficiency and knowledge of social media and various sensitive topics are evenly distributed across groups. If we observe significant differences, we will conduct additional analyses to determine whether they are correlated with other variables and whether they need to be included as control variables in the ANCOVA analysis. We expect to observe a higher accuracy and shorter completion time for the treatment groups to complete the experiment compared with the control group without information cues; furthermore, we expect the highlighting-cue group and label-cue group outperform the probability-cue group. Besides, we expect to see a significant difference between the performance of time-constraint groups and non-time-constraint groups.

5 Contributions

Millions of users uploading large amounts of UGC to popular social media platforms and commercial internet sites all over the world is a conspicuous feature of the contemporary social media landscape. Often owned by publicly-traded firms with a responsibility to shareholders, the mainstream platforms cannot afford any legal, financial and reputational risk that could be caused by UGC (Robert, 2017). Making nuanced decisions on whether each piece of UGC is appropriate currently exceeds the ability of a completely machine-driven or a completely human-driven moderation approach. Therefore, we foresee humans and AI will still need to collaborate with each other in this field in the near future.

Although previous studies have made much progress on exploring both theoretical and practical insights into human-AI collaboration in various scenarios, the knowledge specifically pertaining to the context of content moderation is still absent. What kind of information cues generated by AI could assist human moderators and whether time constraints should be involved to facilitate better performance regarding accuracy and efficiency in content moderation remain unclear. This study is expected to fill the current theoretical gap and contribute to the literature on both human-AI collaboration and content moderation in online social media. We believe moderation strategies based on the answers to our research questions are critical to the reliability of social platforms and endowing users with freedom of speech at an appropriate level.

References

Alkhatib, A., & Bernstein, M. (2019, May). Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, *35*(4), 105-120.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, 110, 486–498.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59-64.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019, October). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (pp. 2-11).

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*, 241–254.

Baron, J. (2012). The point of normative models in judgment and decision making. *Frontiers in Psychology*, 3, 577-578.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems* (pp. 1-14).

Cakmak, M., Srinivasa, S. S., Lee, M. K., Forlizzi, J., & Kiesler, S. (2011, September). Human preferences for robot-human hand-over configurations. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1986-1993). IEEE.

Clark, E., Ross, A. S., Tan, C., Ji, Y., & Smith, N. A. (2018, March). Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces* (pp. 329-340).

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5, 559–583.

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017, May). "Algorithms ruin everything" #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 3163-3174).

Dhar, R., Nowlis, S. M., & Sherman, S. J. (2000). Trying hard or hardly trying: An analysis of context effects in choice. *Journal of Consumer Psychology*, 9(4), 189–200.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4), 1-39.

Dolan, P., Hallsworth, M., Halpern, D., King, D., & Vlaev, I. (2010). MINDSPACE: Influencing behaviour for public policy.

Dudley, J. J., & Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2), 1-37.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. Mcgraw-Hill Book Company.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly*, 45(3), 1527-1556.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. *Information Systems Research*, 33(2), 678-696.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, 28, 210–216.

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-24.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, 112(2), 494–508.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*, 2261–2262.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL: University of Chicago Press.

Hou, J. (2017). Can interface cues nudge modeling of food consumption? Experiments on a food-ordering website. *Journal of Computer-Mediated Communication*, 22(4), 196-214.

Inbar, Y., Botti, S., & Hanko, K. (2011). Decision speed and choice regret: When haste feels like waste. *Journal of Experimental Social Psychology*, 47(3), 533–540.

Johnson, E.J., Shu, S.B., Dellaert, B.G., Fox, C., Goldstein, D.G., Häubl, G., Larrick, R.P., Payne, J.W., Peters, E., Schkade, D. & Wansink, B. (2012). Beyond nudges: Tools of choice architecture. *Marketing Letters*, 23, 487–504.

Kingston, J. K. (2016, December). Artificial intelligence and legal liability. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 269-279). Springer, Cham.

Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition*, 49, 97–122.

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022, April). Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Lam, L., & Suen, S. Y. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(5), 553-568.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision making*. Oxford, UK: Blackwell.

Metzger, M. J., & Flanagin, A. J. (2007). *Digital Media, Youth, and Credibility*. The MIT Press.

Molina, M. D., & Sundar, S. S. (2022). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*, https://doi.org/10.1177/14614448221103534.

Monarch, R. M. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Ordóñez, L. D., Benson III, L., & Pittarello, A. (2015). Time-pressure perception and decision making. *The Wiley Blackwell handbook of judgment and decision making*, *2*, 517-542.

Roberts, S. T. (2017). *Content moderation*. https://escholarship.org/uc/item/7371c1hf.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854-856.

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A user's guide to debiasing. *The Wiley Blackwell handbook of judgment and decision making*, 2, 924-951.

Sonja, S. B. (2014). Evidence-Based Sentencing and the Scientic Rationalization of Discrimination. *Stanford Law Review*, 66(4), 803–872.

Sumita, E., Sugaya, F., & Yamamoto, S. (2005, June). Measuring non-native speakers' proficiency in English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 61-68).

Sundar, S. S, Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME): Four models for explaining how interface features affect user psychology. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 47–86). Wiley Blackwell.

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility* Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, 25(1), 74–88.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. London, UK: Little, Brown.

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13, 1526-1543.

Svenson, O., & Benson, L. (1993). On experimental instructions and the inducement of time pressure behavior. In O. Svenson, & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making*. New York, NY: Plenum Press.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism*, 9(1), 64-83.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295-305).