Integrating Imperfect Machines and Unmindful Users: Assessing Human-Bot Hybrid Designs for Managing Discussions in Online Communities

by

Xinyu Fu

Bachelor of Management, Renmin University of China, 2017

Submitted to the Graduate Faculty of the

Joseph M. Katz Graduate School of Business in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

JOSEPH M. KATZ GRADUATE SCHOOL OF BUSINESS

This thesis was presented

by

Xinyu Fu

It was defended on

April 14, 2023

and approved by

Dr. Priyanga Gunarathne, Assistant Professor of Business Administration, Information Systems and Technology Management

Dr. Laurie J. Kirsch, Professor Emeritus of Business Administration, Information Systems and Technology Management

Dr. Fiona Fui-Hoon Nah, Professor, Department of Media and Communication, City University of Hong Kong

Thesis Co-Advisor: Dr. Dennis F. Galletta, Thomas H. O'Brien Professor of Information Systems, Information Systems and Technology Management

Thesis Co-Advisor: Dr. Narayan Ramasubbu, Professor of Business Administration, Information Systems and Technology Management

Integrating Imperfect Machines and Unmindful Users:
Assessing Human-Bot Hybrid Designs for Managing Discussions in Online Communities

Xinyu Fu, PhD

University of Pittsburgh, 2023

Modern intelligent machines, developed through large-scale data instead of rules, are no longer passive tools waiting to be used, but take proactive actions and work as a co-worker with human agents. They are, admittedly, still imperfect. As a result, machines are often deployed in hybrid "assemblage" configurations that require active human interventions in the workflow. This dissertation examined such a collaborative workflow involving an artificial intelligence (AI) bot and a human to manage inappropriate discussions in the context of online communities. I simulated a bot-assisted news discussion forum where users can affirm or override bot-flagged inappropriate comments. Based on the Information Systems (IS) delegation framework, I examine the main attributes of human agents (cognition), agentic IS artifacts (imperfection), and the delegation mechanisms between these two agents: complacency potential. Cognition, or the "generation effect," occurs when people are asked to explain the bot's activities rather than being told of its flaws. The imperfection has two folds: bots' accuracy and bots' valence. Complacency potential refers to users' tendencies to over-rely on the bot and lack awareness of monitoring the bot's actions. With 1650 subjects over five lab experiments, I found that, on average, users aided by the bot achieved higher decision quality than the ones unaided by the bot. The users that were prompted to provide self-explanation were able to better detect the bot’s errors than others. By deploying the bot at lower accuracy levels, I found that digital platforms may compensate for the lower accuracy of bots by getting users’ active involvement, but there is a threshold level of a bot’s

accuracy below which the bot will not improve the performance of users. Furthermore, the users who encountered a positive bot that recommended good content had higher levels of engagement with the online community relative to those who encountered a negative bot that flagged inappropriate content. Lastly, I found that users who provided a self-generated explanation about bots' actions perceived a higher level of responsibility to monitor the bots' performance, but remained willing to delegate the moderation work to the bot.

Table of Contents

List of Tables

List of Figures

Preface

"If there's a book that you want to read, but it hasn't been written yet, then you must write it."

– Toni Morrison

I would like to express my deepest gratitude and appreciation to all those who have supported me throughout my doctoral journey. Without their unwavering support, this thesis would not have been possible.

First and foremost, I extend my heartfelt gratitude to my advisors, Dr. Dennis Galletta and Dr. Narayan Ramasubbu, for their invaluable guidance, patience, and expertise. Their unwavering commitment to excellence and dedication to my academic and personal growth have been instrumental in shaping the direction and quality of this research. I am truly grateful for their mentorship and the countless hours they have devoted to providing insightful feedback and constructive criticism.

I am immensely thankful to the members of my thesis committee, Dr. Priyanga Gunarathne, Dr. Laurie J. Kirsch, and Dr. Fiona Fui-Hoon Nah for their valuable insights, constructive suggestions, and rigorous examination of my work. Their expertise and scholarly contributions have been instrumental in shaping the intellectual rigor and scientific merit of this research.

I am grateful to the faculty and staff of the Information Systems and Technology Management at Katz Graduate School of Business for providing a stimulating and supportive academic environment. Their commitment to excellence in research and education has been an inspiration throughout my doctoral journey. I would like to express my appreciation to the

administrative staff, Ms. Carrie Woods and Ms. Rachael McAleer, for their assistance and guidance, which has greatly facilitated the administrative aspects of my studies.

My sincere thanks go to my mentors, colleagues, and fellow doctoral students for their camaraderie, intellectual discussions, and unwavering support: Dr. Nanfeng Luo, Dr. Wenxia Zhou, Dr. Yuhui Li, Dr. Yanjun Guan, Dr. Zhen Wang, Dr. Carol Xu, Dr. Qin Weng, Dr. Shadi Janansefat, Dr. Shivendu Pratap Singh, Dr. Di Yuan, Changran Fan, Rachelle Morissette, Joshua Williams, Dr. Yue Zhang, Dr. Peiyuan Huang, Dr. Ankur Arora, Dr. Jisu Cao, Dr. Elizabeth Han, Dr. Xiang Wan, BJ You, Dr. Meizi Zhou, Dr. Mi Zhou, Dr. Yumei He, Dr. Rongrong Zhang, and many others. Their friendship and encouragement have provided solace during challenging times and have made this academic journey an enriching and enjoyable experience.

I am indebted to the research participants who generously contributed their time and knowledge to this study. Their willingness to share their experiences and insights has been invaluable in advancing the understanding of human-AI collaboration.

I am grateful to the funding agencies and organizations that have supported this research financially: Giant Eagle, Inc., and Mr. Ben Fryrear. Their financial assistance has been instrumental in carrying out this study and has allowed me to pursue my academic aspirations.

Lastly, I would like to express my deepest gratitude to my family and friends: my mum and dad, my parents-in-law, my uncle Shijie Chen, Shirley Zhang, Kang Cheng, Wentao Shi, Pengyu Zhao, Yubo Wen, Yuqiao Zhen, Jingjing Yang, Yulun Wu, Jiaqi Guo, Yiwen Zhang, Zhenkun Lian, Yinghua Ding, Dr. Shan Gu, Dr. De Wang, Hongbo Fang, Dr. Xiaomeng Li, Yucheng Wang, Dr. Guangyi Zhao, Tianyu Zhang, and Zhixing Zhang, Anna Lin, Leo Chan, Wong Lee, and all those who have supported me over the years. Their unwavering love, support, and encouragement have been my pillars of strength throughout this journey. Their belief in my

abilities and constant motivation have been instrumental in my success. I am eternally grateful for their patience, understanding, and sacrifices.

I am deeply grateful to my best friend and husband, Sihan Mao, for his unwavering love, support, and encouragement throughout my PhD journey. His respect for me, belief in me, selfless sacrifices, and countless acts of kindness have been instrumental in reaching this milestone. I am truly thankful for his patience and insightful contributions during my dedicated study and writing sessions. Our shared commitment to intellectual growth and dedication to our respective research endeavors have created a nurturing environment that ignites our individual passions. Sihan inspires and motivates me, providing invaluable feedback and thought-provoking perspectives that enrich my work. Our partnership as spouses and research allies fortifies our scholarly pursuits, and I eagerly anticipate the growth and achievements that lie ahead. Sihan, thank you for being my unwavering support system and constant companion on this incredible adventure.

This thesis represents the culmination of years of hard work, dedication, and support from a multitude of individuals. While it is not possible to acknowledge everyone individually, please accept my heartfelt appreciation for all those who have played a role, big or small, in shaping this research. Thank you for your unwavering support, guidance, and belief in me.

1.0 Introduction

In modern society, automation is ubiquitous. To boost efficiency, robots or software bots are increasingly deployed to perform tasks formerly carried out by humans (Davis and Hufnagel 2007, Millman and Hartwick 1987, Parasuraman and Riley 1997, Sandberg et al. 2020). Recent advances in deep learning have shifted the focus of Artificial Intelligence (AI) systems away from static rules toward dynamic learning that is superior at processing unstructured data such as text, image, audio and video (Lecun et al. 2015). As a result, these bots can automate complex problems such as predicting a toxic comment, classifying image content, and analyzing facial expressions (Tarantola 2017). While these bots are generally reliable, they are not error-free and are often deployed in hybrid "assemblage" configurations that involve active human intervention in the workflow (Rai et al. 2019). For instance, in a workflow design the bot may provide a recommendation, which can be subsequently approved or overridden by humans. The bot can provide recommendations based on data and algorithms, while the human can provide context and expertise, ultimately making a more informed decision.

Such expectation of users' active involvement while working with a bot arises from the agentic nature of learning-based systems. As Baird and Maruping (2021) suggested, learning-based systems are no longer passive tools waiting to be used, but rather "agentic." This means that they are no longer always subordinate to the human agent, which leads to ambiguous requirements of the responsibility for tasks. Indeed, for traditional decision support systems, system owners (e.g., companies, developers) are held more responsible for spotting system mistakes than end-users. These systems were created by specialists and are based on explicit standards, so errors are usually foreseeable (Jussupow et al. 2021). In other words, if one error is discovered, it is possible

to address a group of similar errors at the same time. However, users of learning-based systems, which have lately become increasingly popular, are required to engage in the error-catching process; otherwise, they may be more prone to automated mistakes. Learning-based systems' mechanisms are "black boxes," resulting in more unforeseen mistakes (Simester et al. 2020), since they are developed with a data-driven approach. Unlike static traditional decision-support systems, machine learning-based systems continually "learn" through how users interact with them (Jussupow et al. 2021). If end-users are unable to detect faults in learning-based systems, low-quality data will be fed into the loop, allowing the system to "learn" from faulty and even incorrect user input (Neff and Nagy 2016).

Therefore, organizations implementing such bots typically warn users about their potential limitations and seek active user vigilance of bots' actions in order to mitigate overreliance on imperfect bots and avoid serious accidents (Baker 2020, Jussupow et al. 2021, Lambrecht and Tucker 2017). However, it is often the case that such disclaimers are insufficient for triggering desired monitoring behaviors in users (Castelo et al. 2019, You et al. 2022). Prior research has documented that users, even when made aware that algorithms are imperfect, tend to have elevated expectations on their accuracy and heavily rely on them (Galletta et al. 1996, Gunaratne et al. 2018, Logg et al. 2019). At the same time, studies have also revealed that algorithm aversion sets in when users notice even small errors made by bots, and aversive users tend to withhold tasks from bots, which negates any expected performance boosts from AI deployment (Burton et al. 2020, Dietvorst et al. 2015). On average, the decision-making quality of humans who do not rely on assistance from bots is no better than that of a well-developed and finely-tuned algorithm (Dietvorst et al. 2015, Wickens and Dixon 2007). Hence, a nontrivial challenge is to alert users about a bot's imperfections in a way that prompts them to detect erroneous actions of the bot but

without causing algorithmic aversion. This involves managing users' expectations about the bot and designing a human-AI collaborative environment in a way that users do not blindly follow the bot's advice but intervene to override erroneous bot actions, and at the same time evoke users' willingness to leverage the bot's typical strengths in highly scalable data processing, pattern detection, and prediction steps involved in the decision-making process.

To achieve this objective, I followed the Information Systems (IS) delegation theoretical framework (Baird and Maruping 2021) to theorize the human-AI collaborative relationship that improves overall decision quality. I focused on the workflow design where the bot provides a recommendation, which can be subsequently approved or overridden by humans. Based on their 3-step guidance, I examined one relevant attribute of human agent (cognitive capabilities), one main relevant attribute of AI agent (imperfection), as well as foundational mechanisms of delegation.

To boost human agents' cognitive capabilities when working with the bot, I propose to capitalize on the cognitive phenomenon known as the "generation effect" (McCurdy et al. 2020). Explanations generated by individuals through self-explanation is known to be better retained and more easily retrieved than information that is simply presented to the individuals. This is because generating an explanation requires individuals to actively engage with the context, which triggers a deeper level of processing and promotes the ability to make more connections between different pieces of information gathered from the context (Hitron et al. 2019, Keil 2006). In contrast, when individuals are directly presented with an insight or explanation, they may be more passive in their processing of the material and may not engage with it deeply (Metcalfe 2017).

A human-AI collaboration design that takes advantage of the generation effect would prompt users to generate their own explanations of a bot's actions and more actively discover any

errors made by the bot (Keith and Frese 2008). I propose that such a design would be more effective than commonly practiced procedural training approaches that directly inform users about bot imperfections. When users generate self-explanations about the bot's actions, such as asking "please provide an explanation about how the bot may come to this suggestion," they would realize that the task at hand is challenging and are more likely to develop a sympathetic attitude towards the bot, which would serve to counteract any negative emotions that trigger algorithm aversion. In addition, users who better understand the bot's imperfections and working mechanisms would continue to collaborate with the bot, as they would be motivated to rectify errors and help the underlying AI system to learn from the user-generated fixes. Thus, a human-AI collaborative task design that leverages the generation effect can be expected to better engage users for error detection and contribute to higher decision-making quality.

In addition, I am also interested in testing whether the effect of self-generated explanations could endure over multiple tasks. Individuals' cognitive resources may deplete when engaging with more tasks and thus, the benefits of self-generated explanation may diminish as users progress through multiple tasks (Sternberg 2000). Specifically, I seek to determine whether the benefits of self-generated explanations can spill over to subsequent tasks, or whether they are limited to the immediate task following the self-explanation prompt.

From the perspective of an AI agent, I am interested in examining how the impact of the generation effect on decision-making varies with the level of a bot's imperfection (Hardin et al. 2018, Schuetzler et al. 2020). Previous research on automation-induced complacency has suggested that reducing the accuracy level of an automated tool may be an effective way to reduce a decision maker's complacency, as users may become hesitant to delegate work to a bot that is perceived as inaccurate (Parasuraman and Manzey 2010). With a reduced willingness to alleviate

workload to the less accurate bot, it remains unclear, however, whether users will realize a higher level of error detection and improve performance. In fact, as recently reported evidence shows, users may completely lose interest in working with algorithms perceived as inaccurate and their willingness to monitor the bot's performance would wane (Burton et al. 2020, Dietvorst et al. 2015). I aim to explore whether users who are prompted to provide a self-explanation may react differently when working with a bot at various levels of accuracy.

Furthermore, the valence of bots, which refers to their inherent positive or negative qualities, is another important factor to consider when designing and implementing automated systems. Depending on the type of task or function assigned to a bot, its valence can represent various types of imperfections that may affect its performance. For example, a bot with a positive valence may exhibit a tendency to be overly optimistic or take unnecessary risks, while a bot with a negative valence may be overly cautious or hesitant (Han et al. 2022, Zhang et al. 2010). It is essential to understand the potential imperfections associated with bots' valence, as this knowledge can inform decisions about the appropriate level of human oversight and intervention necessary to ensure optimal performance and outcomes.

Finally, for exploring a tangible pathway through which the generation effect transpires to positively impact decision-making quality in human-AI collaborative tasks, I focused on Automation-induced complacency potential (Chan Fung 2021, Moray et al. 2000). In the human-AI collaboration context, complacency is an unjustified assumption about the AI system, which may result in non-vigilance of the system’s errors (Singh et al. 1993). Specifically, complacency has been measured through two dimensions in prior research (Merritt et al. 2019): (a) Alleviating Workload: the attitude about using automation to ease workloads, and (b) Monitoring: the lack of degree of attention devoted to monitoring automated tasks. I propose that a human-AI

collaboration design that leverages self-generated explanations reduces automation-induced complacency potential, which, in turn, serves to improve overall decision-making quality.

In summary, the purpose of this dissertation is to: (1) examine the effects of self-generated explanation on decision-making quality in a human-AI joint decision-making task; (2) explore whether the effects of explanation on decision-making quality will change with different levels of the bot's accuracy and different bots' valence; (3) explore the role of complacency as an underlying mechanism. I proposed to test these questions in a content moderation task. The content moderation task is particularly suitable for my study for the following reasons: First, assessing comments in a discussion forum is increasingly becoming a human-AI collaborative endeavor in many settings. For example, the New York Times, the Washington Post, and the Wall Street Journal have trained machine-learning algorithms with millions of labeled comments and get human users involved for the assessment of bot's actions and edge cases (Marvin 2019). Bots implementing such machine-learning algorithms do make errors, and organizations deploying them desire to get feedback from human users for continuously training and improving the algorithms (Fügener et al. 2021a, b). Thus, the content moderation task provides a timely and relevant real-world context for our experiments. Second, assessing comments in online discussion forums is a general and typical real-world task that does not require any specialized training, as it has been reported that almost sixty percent of the worldwide population, or 4.76 billion of the 8.01 billion people on the planet actively visit online communities (Kemp 2023). Finally, I chose the online discussion forum context to minimize users' concerns about reporting errors. Within organizations, individuals may be reluctant to report errors, for example to avoid conflicts (Keith and Frese 2008, Zhao and Olivera 2006). Hence, I chose a context free from structural factors that may prevent individuals from detecting and reporting. In the next section, I discuss the practical

background of the content moderation, and why human-AI hybrid design can be a potential approach to managing online comments.

1.1 Challenges of Managing Online Discussions

Online communities have risen in popularity from the realm of computer hobbyist affinity groups, to society in general, and finally, to firms (Bapna et al. 2019, Ransbotham and Kane 2011). Organizations can benefit from online communities in multiple ways, such as acquiring new customers (Trusov et al. 2009), creating a network effect (Oestreicher-singer and Zalmanson 2013, Shriver et al. 2013), promoting information sharing (Pavlou and Dimoka 2006), improving companies' understanding of customers and markets (Urban and Hauser 2004), and ultimately boosting sales of products and related customer expenditures (Clemons et al. 2006, Das and Chen 2007). Given these advantages, operating online communities and managing user-generated content have become well-accepted as an important part of a digital platform's strategy. The success of online communities depends on their members' continued contribution of high-quality content, as well as the proper and timely moderation of this user-generated content to ensure the community's long-term viability (Chen et al. 2018, Faraj et al. 2015, Johnson et al. 2014). Otherwise, inappropriate behaviors (e.g., trolling, cyberbullying) across various online communities will threaten users' engagement (Barlett 2017, Gu et al. 2007).

Nevertheless, the cost of managing the quality of a high volume of user-generated content is economically high for organizations and psychologically high for moderators (Feldman 2019) . For example, Facebook, with a worldwide reach of about 2.3 billion users, employs approximately 15,000 moderators at $28,800 each (Feldman 2019) to identify and remove violent, sexually

explicit, and offensive content on the Internet (Thomas 2020). Similarly, the Guardian, which receives around twenty thousand comments every day, hires moderators to review and remove comments violating community rules; from 2010 to 2016, about 1.4 million noncompliant comments have been discarded by these moderators (Gardiner et al. 2016). The high cost of moderating user-generated content has prompted some platforms to suspend readers' ability to comment on sensitive or controversial news items (Etim 2017), or like CNN.com, to shut down the online discussion feature altogether (Gross 2014).

To alleviate the limits imposed by the restricted number of professional moderators, platforms tried crowdsourcing moderation duties to users. Unfortunately, such efforts often end in failure. For example, Civil Comments created a peer-review submission process to replicate face-to-face social interactions, where commenters are required to score the civility of three randomly selected remarks before their own is rated by others (Bogdanoff 2015). The notion has sparked widespread attention, since users' self-governance is supposed to empower people to establish the community they desire (Lomas 2015). After three years, however, the platform's owners decided to shut it down, because community members downvoted comments with opposing opinions and the self-organizing community resulted in anarchy (Bogdanoff 2015). In short, the failure of Civil Comments demonstrates relying on users, at least solely relying on users, is not a feasible strategy.

Another approach to helping reduce the workload of moderators and the associated costs is to automate the moderation process, especially using machine learning-based algorithms (i.e., bots) for moderation tasks (Tarantola 2017). The prediction technologies used by some moderation bots, using deep learning algorithms in conjunction with natural language processing (NLP), tend to have high accuracy, but for the foreseeable future they are unlikely to yield perfect classifications of real-world content into appropriate and inappropriate categories. Furthermore,

they have little transparency and, thus, have low predictability of errors (Jussupow et al. 2021). As a result, bots are often deployed in hybrid "assemblage" configurations that involve active human intervention in the content moderation workflow (Rai et al. 2019). For example, the New York Times, the Washington Post, and the Wall Street Journal reportedly train bots with millions of labeled comments and deploy these bots as decision aids for professional content moderators (Marvin 2019, Renner 2016). Retaining professional moderators in the workflow, however, limits the scalability of community deployment.

Table 1 Approaches to Moderating Online Discussions

	Professional moderators	Crowdsourced to users	Machine
Professional moderators	Moderators only: Costly; Limited Scalability		
Crowdsourced to users	Moderator + Users: Low user involvement	Users only: Users may not be accountable	
Machine	Moderator + Machine: Still suffers from limited scalability	Uesrs + Machine: User-Bot Hybrid Design: Current dissertation	Machine only: Machines are imperfect

The real-world applications showed that relying entirely on users or bots is ineffective. Additionally, assemblage configurations of professional moderators and machines, or assemblage configurations of professional moderators and users, continue to suffer from the problem of constrained scalability (See Table 1 for a summary). However, to the best of the authors' knowledge, no study has been conducted on a hybrid user-machine approach. Specifically, platforms can deploy content moderation bots and depend on the judgments of end-users in the audience to verify the accuracy of the bot's action. Such a crowdsourced approach to achieve human-AI assemblage for content moderation is scalable and assists in continuous improvement of the performance of the bot.

1.2 Assessing a Human-Bot Hybrid Design of Managing Online Discussions with IS Delegation Framework

To explore the possibility of a human-bot hybrid design for content moderation, I adopted the IS delegation theoretical framework (Baird and Maruping 2021) to guide my designs. The IS delegation framework provides guidance to examine the relationship between human agent and agentic IS artifacts. A key characteristic of such agentic IS artifacts is that they are not passive tools waiting to be used, but can take proactive actions and not always subordinate to the human agent. Thus, when assessing a hybrid human-bot hybrid design, such a framework emphasizes the contribution from both agentic IS artifacts and human agents, but does not give primacy to human agency. The IS delegation framework provides a new perspective to examine the responsibility for tasks with ambiguous requirements between the human and agentic IS artifacts.

Based on their guidelines, the first step is to identify and explicate the salient attributes of the task or desired outcome under study. In a crowdsourced, hybrid human-AI configuration for content moderation, there are two sets of desired outcomes. The first one is comment-moderation quality. In the hybrid human-AI design, end users are expected to interpret the bot's action and verify its validity. It is, however, not clear that individual users can successfully augment their decision-making with bot advice (Burton et al. 2020, Jussupow et al. 2021). On the one hand, bots may provide valuable assistance to human moderators to quickly sift through voluminous content and identify inappropriate comments. On the other hand, evidence from prior literature highlights that users tend to routinely follow automated decision aids and overlook errors made by those decision aids even if the errors are obvious (Alberdi et al. 2004, Galletta et al. 1996, Maltz and Shinar 2004, Metzger and Parasuraman 2005). Thus, benefits from imperfect decision aids may be hard to realize because users often fail to reject incorrect system advice and tend to be misled

by them (Chan et al. 2017, Heart et al. 2011, Wickens and Dixon 2007). In addition, users may become averse to advice provided by the bot once they are aware of the bot's erroneous actions, even as the bot's algorithm learns from those errors (Dietvorst et al. 2015). The second desired outcome is engagement. It is an open empirical question how user engagement in online communities change in the presence of such a highly accurate but still imperfect bot.

The second step is to identify and analyze salient delegation mechanisms relative to the task under study. As discussed in the previous section, a nontrivial challenge in hybrid human-AI design is to alert users about a bot's imperfections in a way that prompts them to detect erroneous actions of the bot but without causing algorithmic aversion. Thus, the delegation mechanism of interest here is to understand users' willingness to delegate tasks to the bot as well as their monitoring behaviors. Accordingly, I identified the salient attributes of agents relative to the task and delegation mechanisms. The desired contribution from a human agent in a hybrid design is their cognitive capabilities, i.e., their understanding about the bot's actions. As mentioned earlier, IT artifacts like imperfect bots can make end users vulnerable to simplification, following rules even when exceptions are needed (Butler and Gray 2006, Jussupow et al. 2021, Valorinta 2009). In order to improve the decision-making performance of users, it is crucial to improve the awareness of users about the imperfections of the decision aids and the need to systematically monitor their actions. Ackerman and Thompson (2017) propose that users can better monitor and control their decision-making using metacognition. When decision-making end users are aided by an imperfect bot, users need to not only monitor the progress of their own reasoning and problem solving but also the performance of the imperfect decision aid (Jussupow et al. 2021). By being actively involved in the task, end users can increase their awareness of the need to jointly monitor and regulate their own task activities along with the bot's actions. When working with an imperfect

bot, prompting users to provide a self-explanation has the potential to mitigate users' tendencies to blindly follow the bot's advice and avoid being misled by the bot's erroneous actions.

I propose that end users can get actively involved by using the self-generated explanation technique. By pausing and explaining their observations, individuals uncover the underlying structure of task actions and acquire the knowledge for predicting and controlling their future actions (Keil 2006, Lombrozo 2006, Lombrozo and Carey 2006). When individuals provide explanations—even to themselves—they trigger metacognitive monitoring, process information more effectively, are more able to generalize what they have learned, and, therefore, are more readily able to apply their knowledge to novel situations. This phenomenon is known as the generation effect in cognitive psychology (McCurdy et al. 2020, Nass et al. 1994, Williams and Lombrozo 2013). To get individuals engaged with deeper thinking and exploration after encountering evidence that challenges their "assumptions" about the accuracy of the bot (Liquin and Lombrozo 2017), the active involvement triggered by self-explanation may improve users' attentiveness to the bot's actions and helps them better detect the bot's errors. Building on this, I propose that prompting end users who are engaged in online content moderation to explain the imperfect bot's actions has the potential to improve their comprehension level of bots' imperfections, which, in turn, helps them to detect the erroneous actions of the bot and improve task performance. Thus, I seek to empirically test the presence of the self-generated explanation effect in the context of bot-aided content moderation through my first research question:

RQ1: Will participants of an online community who utilize an imperfect bot for content moderation and are prompted to explain the bot's actions achieve higher moderation quality and better engagement?

Since cognitive resources are limited and need to be allocated depending on task priorities, the effects of self-generated explanation might wear off as users progress through multiple tasks (Hollender et al. 2010). Users who are prompted to provide explanation at the beginning of their content moderation task pay increased attention to the decision-making context, but they may also be prone to depleting their cognitive resources (Sternberg 2000). After users complete a task immediately following their self-explanation, it is uncertain whether they will still have enough mental resources to continue the deliberate regulation of metacognition and effectively complete follow up tasks. Once a user's mental resources are exhausted, their capacity to exercise self-control declines (Baumeister 2018, Schmeichel et al. 2003). As a result, the user's original system monitoring state might diminish as they embark on subsequent tasks. I examine the durability of the self-explanation effect through the following research question:

RQ2: Will the effects of self-explanation endure multiple tasks?

The salient characteristic of the bot in this dissertation is its imperfection. A bot's inherent accuracy represents the level of imperfection (Hardin et al. 2018, Schuetzler et al. 2020). Previous research has suggested that reducing automation's accuracy level may be an effective way to reduce a decision maker's complacency, as users may become hesitant to delegate work to a bot that is perceived as inaccurate (Parasuraman and Manzey 2010). With a reduced willingness to allocate workload to the less accurate bot, it remains unclear, however, whether users will realize a higher level of error detection and improve performance. In fact, as recently reported evidence shows, users may completely lose interest in working with algorithms perceived as inaccurate and their willingness to monitor the bot's performance would wane (Burton et al. 2020, Dietvorst et al. 2015). I aim to explore whether users who are prompted to provide a self-explanation may react

differently when working with a bot at various levels of accuracy. Thus, my next research question is:

RQ3: What is the impact of the accuracy of the bot on users' moderation decision quality and engagement?

Studies in positive psychology have shown that positive and negative orientation can have differential influences on individuals' performance. A moderation bot can either be designed to alert users to the presence of inappropriate content (i.e., negative bot) or to compliment or recommend to others by "upvoting" insightful content (i.e., positive bot). A positive orientation is expected to increase individuals' psychological well-being, joy, and engagement, whereas a negative orientation (e.g., strong censorship for violating a norm) might be more potent in improving individuals' content moderation task performance (Apter 2018, Ilgen et al. 1979). In the context of an online community platform, positive or negative orientation is a crucial design choice, and real-world applications vary in their valence orientation. For example, the New York Times designed its discussion forum to provide positive feedback by drawing attention to insightful comments through "Editors' Picks." On the other hand, Twitter, Facebook, and Zhihu, the largest Q&A community in China, tag suspicious posts with a warning. I seek to examine the differential effects of the valence of the bot's processing orientation: positive (highlighting insightful comments), or negative (tagging inappropriate comments). Thus, my next research question is:

RQ4: What is the impact of the valence of the bot's processing orientation (positive versus negative) on users' engagement?

Finally, I explored the mechanism underlying the effect of self-explanation on decision making quality. Automation-induced complacency, or sub-optimal monitoring of automation

performance, has been cited as a contributing factor in numerous major transportation and medical incidents (Chan Fung 2021, Merritt et al. 2019). Research has widely discussed when and why complacency occurs (Metzger and Parasuraman 2005, Parasuraman and Manzey 2010, Singh et al. 1993). For example, researchers have examined how individuals with a greater inclination toward complacency than others (i.e., complacency potential) have a lower error detection rate (Farrell and Lewandowsky 2000, Merritt et al. 2019). Self-explanation can trigger individuals to engage with deeper level information processing and enable them to process information more effectively. Thus, my last research question is:

RQ5: Is reducing complacency the underlying mechanism of the impact of self-explanation effect on users' moderation decision quality and engagement?

I summarized my theoretical framework in Figure 1. To achieve the desired outcomes of decision quality and engagement, I proposed to examine the cognitive actions as human agents' attributes, and imperfection as AI's attributes. Cognitive actions are actions involving comprehension, such as building and revising mental models about how the bot functions as well as knowledge retention and expansion. Thus, in this study, cognitive actions refer to processing of the bot's function: whether their comprehension level of the bot has a higher level due to self-generated explanations, and how long such effects could endure. The imperfection has two folds in our study: the first one is accuracy, which refers to level of imperfection; the second one is valence, which refers to the type of imperfection (e.g., making an incorrect assessment about recommending an insightful comment vs. making an incorrect assessment about detecting an inappropriate comment). I propose that the effective delegation in such a hybrid human-AI design is through reducing individuals' complacency potential. In other words, individuals who are prompted to provide a self-generated explanation about bots' actions will have increased

awareness to monitor the bot agents' performance, but will remain willing to delegate tasks to the bot.

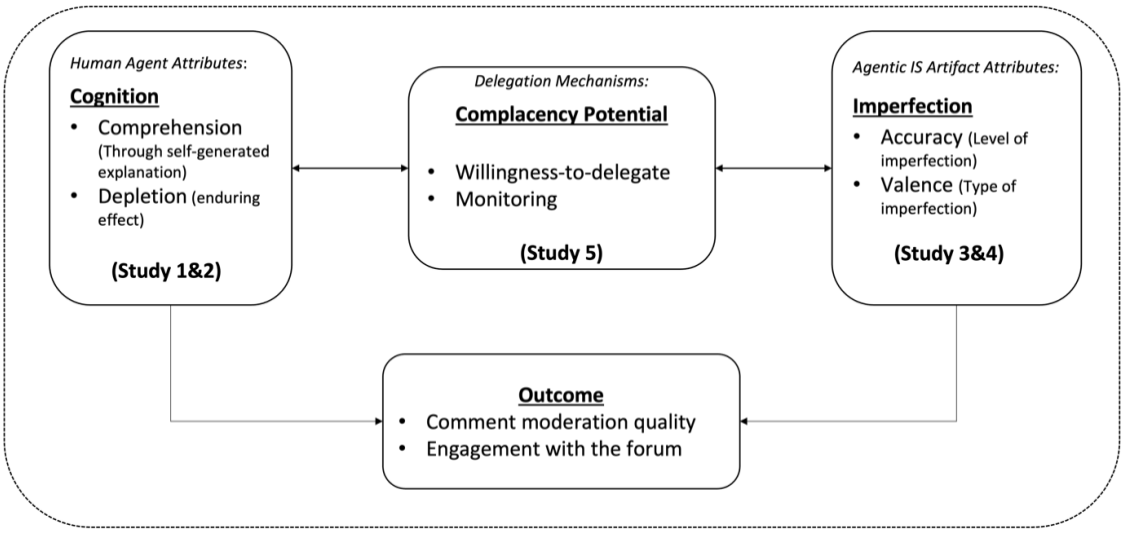


Figure 1 Overarching Framework

To answer these research questions, I collected data through five waves of randomized control trials that utilized a simulated news discussion forum and offered an imperfect bot with and without self-explanation prompts, positive or negative valence, and different accuracy levels as distinct treatment conditions for users. The analysis revealed that, on average, users in the self-explanation treatment condition were able to better detect the bots' errors and achieve higher performance than those directly informed about bots' imperfections, but the benefits wore off over subsequent tasks. Those who were directly informed about bots' imperfections became aversive and tended to discard the bots' suggestions. In contrast, users who actively discovered the bots' imperfections through generating an explanation about bots' actions had increased awareness to monitor the bot's performance and still had willingness to work with the bot. Moreover, users aided by a positive bot had higher levels of engagement with the online community relative to

those who encountered a negative bot. I also found that when bots performed at a low accuracy level (between 50%-60% accuracy), users lost interest in working with the bot and did not gain performance benefits when using the bot. Thus, the results suggest that deploying highly accurate though imperfect bots in digital platforms can be beneficial; Providing self-explanation could help users to become more vigilant to the actions of imperfect bots; and there are tradeoffs between human-bot hybrid designs that encourage engagement and those that emphasize moderation decision quality.

1.3 Literature review and hypothesis development

1.3.1 Detecting Algorithmic Errors in Human-bot Hybrid Designs

In human-AI collaborative tasks, errors may go undetected simply because individuals were not expecting them to happen in the first place. Research shows that users have uncritical reliance on algorithmic accuracy and sometimes may even prefer advice generated by algorithms (Gunaratne et al. 2018). Users may simply follow the advice given by an algorithm even when it is wrong (McKnight et al. 2020, You et al. 2022). When users have uncritical reliance on algorithms, the consequences can, however, be severe (Mahlfeld et al. 2011, Skitka et al. 2009). For instance, there have been reports of serious accidents due to pilots' overreliance on aircraft automation and medical errors that can be attributed to care givers' failures to anticipate and detect automation errors (Berton 2018, Degani 2003, FDA 2022, Institute for Safe Medication Practices 2017). In the context of decision-making, various studies show that users' performance may be degraded if they blindly follow suggestions provided by the algorithms (Galletta et al. 2005).

In contrast to overreliance, humans interacting with algorithms may also have negative cognitive and behavioral reactions and develop strong aversion to algorithms (Naquin and Kurtzberg 2004, Park et al. 2022, Srinivasan and Len Sarial-Abi 2021). One possible explanation for this phenomenon is that users tend to hold algorithms to a high standard of perfection, but when they encounter algorithmic errors and the algorithms fail to meet the high expectations, users become disappointed, and aversion sets in (Burton et al. 2020). When users disregard suggestions from a collaborating bot due to algorithm aversion, their overall decision-making quality is likely to suffer because users fail to take advantage of the data processing strengths of the bots, but also miss the opportunity to refine the bots by detecting and fixing their errors.

The issues related to both overreliance and aversion of algorithms become particularly pronounced in the context of bots developed using modern machine-learning approaches. Learning-based algorithms are trained using a large sample of observations that include both structured and unstructured data such as text, image, audio and video, and the training is often without any specific domain-based or expert-defined rules. For example, in the context of content moderation, bots based on machine-learning algorithms can be trained using historical data that has classified comments into appropriate and inappropriate categories based on input from end users and professional moderators. Even after initial deployment, the bots can be enabled to continuously learn from end-user actions and feedback. Such continuous learning-based systems, however, have been termed “black box” systems because the underlying mechanisms through which they predict and formulate actions are complex and cannot be described with deterministic rules (Simester et al. 2020). Thus, the patterns of errors made by these systems are often unpredictable and the root causes of the errors are difficult to detect and explain (Jussupow et al. 2021, Simester et al. 2020). In this context, while it is difficult for users to anticipate and make

sense of the errors committed by bots, users can contribute to continuous improvement by actively noticing and reporting the errors, which will be subsequently used to retrain the learning-based models. If end users fail to detect and report errors, low-quality data will be fed into a vicious loop of training, resulting in further deterioration of the learning-based systems (Neff and Nagy 2016). Hence, active user participation in error detection and rectification is essential for continuous improvement of bots based on machine-learning algorithms.

Efforts have been made to address the challenges associated with detecting bot errors by providing users with educational materials to enhance users' comprehension of the algorithms. A common approach is to display warning messages about a bots' imperfections (Baker 2020, Lambrecht and Tucker 2017). Experiment results show that, however, a disclaimer about the imperfections does not sufficiently increase error detection behaviors (Castelo et al. 2019, You et al. 2022). Another approach is to increase algorithmic transparency and help improve a user's understanding of the mechanisms implemented through the bots. For example, prior research examined the impact of offering detailed explanations about how a system is designed to enhance the initial trust of users in recommendation algorithms, but it is unclear whether the improvement in users' trust in the algorithms would increase the likelihood of detecting algorithmic errors (Wang and Benbasat 2014). Finally, prior studies have tested the role of increasing user exposure to a variety of real-world algorithmic errors on post-exposure task performance (Bahner et al. 2008, Dietvorst et al. 2015, Skitka et al. 2009). Exposure to algorithmic errors allows users to better calibrate their expectations about the algorithms, which can help improve error detection and collaborative performance, but as mentioned before, a higher level of exposure to algorithmic errors may also trigger algorithm aversion (Burton et al. 2020, Dietvorst et al. 2015). Overall, prior approaches have focused on presenting users with information with the goals of improving

algorithm transparency and increasing users' understanding of the algorithms. In prior approaches, however, it is not clear if users have moved beyond being passive receivers of helpful information and if they are deeply engaged with the information to enable active learning (Montazemi and Wang 2015). I address this gap in this study and examine how active learning can be triggered and how that may aid in proactive discovery of bot errors and improvement in decision quality.

1.3.2 Human Agents' Cognitive Capabilities

Cognitive psychologists have been studying human mental processes, such as attention, perception, memory, and problem-solving, for decades (Boag et al. 2019). The primary objective of this line of research is to understand how individuals acquire, process, store, and retrieve information (Evans 2008). While AI bots excel at processing large-scale, similar information, they may struggle with corner cases or less common scenarios with insufficient data from which to learn (Simester et al. 2020). Therefore, it is crucial to explore the cognitive capabilities of human agents in complementing AI bot agents by correcting their errors. One area of interest in cognitive psychology that could improve comprehension is the generation effect, a phenomenon where actively generating information leads to better memory retention and knowledge comprehension than simply reading or listening to information (Slamecka and Graf 1978). I propose that prompting individuals to provide self-generated explanations about bots' actions can increase their cognitive understanding of bots' imperfections, and thus benefit subsequent error detection behaviors. Another relevant concept in cognitive psychology is the depletion of cognitive resources, which helps us understand whether the effects of self-generated explanations can endure (Muraven et al. 2019). Cognition is a limited resource that can become depleted with prolonged use, leading to diminished cognitive ability to process information in subsequent tasks (Baumeister

2018). Next, I reviewed key literature on these concepts to illustrate how human agents process information and how these processes can influence their cognitive abilities when interacting with a bot.

1.3.2.1 Information Comprehension Through Self-Generated Explanations

One strategy for fostering active user engagement with errors is to encourage self-generated knowledge (Keith and Frese 2008, Metcalfe 2017). The Cognitive Psychology literature indicates that training processes that provide learners with the opportunity to generate knowledge, rather than passively receive information, strengthen comprehension and memory (McCurdy et al. 2020). Such a strategy has been known to elicit deeper cognitive processing and increased understanding, and the generation effect has been demonstrated in diverse training contexts (Hinojosa et al. 2017, Hitron et al. 2019, Jensen et al. 2017, McCurdy et al. 2020).

I propose leveraging the generation effect by prompting users working with bots to provide self-generated explanations about the bots' actions. An explanation is a statement that satisfies a user's need for reasoning that can differentiate the occurrence of an observation that requires explanation (i.e., the explanandum) from the contrasting nonoccurrence of alternative events (Horne et al. 2019, Khatri et al. 2018). As an illustration, suppose a bot labels a comment in a discussion forum as inappropriate. A potential (but unobserved) alternative to this action would be marking the comment as appropriate. By providing self-generated explanations about the bot's actions, users can either identify reasons why the comment did not comply with the forum's commenting rules, or challenge the bot's decision by suggesting that it committed an error in this particular case.

When prompted to provide an explanation about a bot's actions, users must invest additional effort (i.e., deliberate thinking), for example, conduct what-if analysis about the

observed and potential alternative scenarios, rather than readily accepting the bot's actions through automatic thinking processes (Khatri et al. 2018). By engaging in the process of generating explanations for the bot's actions, users may gain a deeper understanding of the system's functioning and identify areas for improvement. Such a triggering of deliberate cognition by seeking an explanation would, in turn, aid beneficial metacognition, i.e., "the thinking about thinking" processes that regulate decision making (Ackerman and Thompson 2017). In the context of decision making with assistance from a bot, metacognition includes monitoring the performance of both the decision makers themselves and the bot (Jussupow et al. 2021). By developing an explanation for the bot's actions, users improve their awareness of the bot's imperfections and the need to monitor these imperfections' consequences.

Furthermore, self-generated knowledge about a bot's actions can also mitigate any negative reactions towards algorithmic errors. By attempting to explain how the bot formulates its assessments, users may gain a sense of the challenges involved in the task, both for humans and for bots. Consequently, users who generate explanations for the bot's actions tend to be more understanding of algorithmic errors and may view their occurrences as opportunities to contribute towards further system improvement. Previous research on error management training has highlighted the benefits of creating an atmosphere in which errors are better understood, and studies have emphasized how informational feedback about errors benefits user learning and engagement (Alfieri et al. 2011, Keith and Frese 2008). In contrast to training methods that rely on direct instructions, an approach that facilitates self-generation of explanations moves users beyond being passive bystanders and encourages active user involvement (Canning and Harackiewicz 2015, Hitron et al. 2019). Thus, generating self-explanations about a bot's actions

represents an active discovery approach for users, and I posit that it would improve error detection and decision-making quality in human-bot collaborative tasks. This is my first hypothesis:

H1(a): When working with a bot, users who provide self-generated explanations about the bot's actions will have higher decision-making quality, relative to other users who do not provide such explanations.

When users moderating content are prompted to explain a bot's actions, and in turn, increase their comprehension of bots' imperfections, they spend more cognitive resources processing the information, relative to other users who are not subject to the explanation treatment. I propose that a discussion forum user who provides a self-generated explanation is engaged in a deeper level of information processing, and therefore, has more opportunities to realize the need for active engagement. Engagement in online communities can be defined as the level of interaction, involvement, and participation individuals have in the community's activities and discussions. It refers to the extent to which community members actively contribute to the community's shared goals, values, and interests. Some key indicators of engagement in online communities include the frequency and quality of contributions, such as posting, commenting, sharing, and liking. Other metrics include the number of views, replies, and shares that community content receives, as well as the level of social interaction among members, such as private messaging and group discussions. Users doing the self-explanation exercises tend to allocate more effort and time to understanding the rules of the community and managing bots' imperfections, which increases their overall sense of responsibility to contribute to the community's goal of maintaining a healthy discussion environment. Relegating the responsibility to an imperfect bot would cause cognitive dissonance in such users (Festinger 1962, Joyce and Kraut 2006), which

they tend to avoid by maintaining higher levels of engagement with the community. Thus, my next hypothesis is:

H1(b): When working with a bot, users who provide self-generated explanations about the bot's actions will have higher engagement, relative to other users who do not provide such explanations.

1.3.2.2 The Depletion of Comprehension over Tasks

Individuals' cognitive resources are limited, and thus, they need to allocate attention based on task priority (Hollender et al. 2010). After allocating effort and attention to a certain task that is of importance, the amount of available cognitive resources decreases. As suggested in Ego depletion theory (EDT), possessing cognitive resources that can become depleted after prolonged use leads to decreased ability to exert self-control in subsequent tasks (Muraven et al. 2019). The concept of ego depletion has been widely studied and debated in the field of psychology, especially in the domain of decision-making (Inzlicht and Friese 2019).

The initial studies on ego depletion theory conducted by Roy Baumeister et al. in the late 1990s showed that participants who engaged in a task that required self-control, such as resisting eating chocolate or suppressing emotional reactions, performed worse on a subsequent task that also required self-control, such as solving anagrams or persisting on a difficult puzzle (Baumeister 2018). These studies provided support for the idea that self-control is a limited resource that can become depleted.

However, subsequent studies have produced mixed results, with some replicating the original findings and others failing to do so (Inzlicht and Friese 2019). Some researchers have suggested that the initial studies suffered from methodological flaws such as small sample sizes, lack of control conditions, and publication bias. Other researchers have suggested that the effect

of ego depletion may be influenced by individual differences such as motivation, belief in willpower, and trait self-control, or other factors such as glucose levels, sleep deprivation, and social support (Baumeister et al. 2006, Muraven et al. 2019). A recent meta-analysis conducted by Hagger et al. (2010) found a small to moderate effect size for ego depletion on subsequent self-control tasks, but also noted that the effect was highly dependent on the type of self-control task used and the interval between the tasks.

The typical depletion tasks in the decision-making domain rely on consuming cognitive resources by inserting high-level cognitive dissonance (Muraven et al. 2019). A widely adopted task is the Stroop task, which requires individuals to name the color of the ink in which a word is printed, while ignoring the word's meaning (Johns and Butler 1991). This task is difficult because the word's meaning often interferes with the individual's ability to name the color, and thus it requires a high degree of self-control. Another typical task is The Go/No-Go task, which requires individuals to respond quickly to a stimulus (e.g., a green light), but inhibit their response to another stimulus (e.g., a red light) (Gomez et al. 2007). This task requires the individual to constantly monitor their responses and resist the impulse to respond to the wrong stimulus. However, such tasks are less prevalent in real-world settings and are typically unrelated to the primary task at hand. There have been recent calls to examine cognitive depletion using tasks that occur in real-world settings (Hagger et al. 2010). For instance, in my study, content moderation can be a task that necessitates significant cognitive resources and effort. Having individuals complete multiple sessions of the content moderation task provides an opportunity to observe the depletion of their cognitive resources. I am interested in investigating whether the benefits of self-generated explanation can endure multiple tasks.

In sum, each time users engage in monitoring behavior, which is effortful and self-regulatory, their cognitive resources are depleted, and subsequent moderation decision quality may suffer. Users have limited mental resources to regulate and persist at particular behaviors. As asserted in EDT, all acts of self-regulation draw on a central pool of self-regulatory resources, and when individuals have already consumed some self-regulatory resources, they need time to recover (Muraven and Baumeister 2000). The theory specifies that when individuals are in a state of ego depletion as a result of prior self-control exertions, their capacity to exercise subsequent self-control declines. Empirically, ego depletion has been found to influence a wide array of behaviors that require self-control resources such as logic and reasoning (Schmeichel et al. 2003), information processing (Fischer et al. 2008), and supervision (Barnes et al. 2015, McAllister et al. 2018). Because moderation tasks involve cognitive resource consumption activities including logic and reasoning, information processing, and supervision, I posit that deliberate use of cognitive resources in monitoring system performance may not endure multiple tasks. In a content moderation context, paying attention to a bot's errors during initial moderation tasks would deplete actors' self-regulatory resources (Muraven 2012), which is consequential for engagement and moderation decision quality in later tasks. Therefore, my next hypothesis is:

H2: The positive effect on a user's moderation decision quality resulting from the prompt to explain the actions of the bot declines over multiple tasks.

1.3.3 AI bots' Imperfection

1.3.3.1 The Algorithm Accuracy Effect

Accuracy, a common metric for assessing algorithm performance and imperfection levels, is measured as the percentage of accurately assessed cases among all cases. This metric provides

users with an indication of an algorithm's reliability (Metzger and Parasuraman 2005). In general, research shows that people are more likely to trust and rely on accurate algorithms than inaccurate ones (Atanasov et al. 2020, Eric and John 1986, Yin et al. 2019). However, users will stop engaging with a bot once their trust in it falls below a certain threshold (Hardin et al. 2018, McKnight et al. 2020, Schuetzler et al. 2020). A meta-analysis of decision support systems has shown that users tend to find automated tools helpful when the algorithm's accuracy is at or above 70% (Wickens and Dixon 2007). This threshold was confirmed by a recent study using randomized lab experiments to test the effect of algorithmic accuracy on human trust and system usage (Yu et al. 2019). If the accuracy falls below this threshold, task performance may be worse than if no tools had been used at all, as users may reject suggestions from these automation tools and rely solely on their own judgment (Castelo et al. 2019, Wickens and Dixon 2007).

A significant body of research has investigated the behavioral implications of algorithm accuracy, with a focus on how this information is communicated to users. Accuracy is typically communicated to users through either direct verbal descriptions that summarize algorithmic performance, such as expressing the diagnostic accuracy of a medical algorithm as 87% (Longoni et al. 2019), or through multiple rounds of interactions and feedback (Dietvorst et al. 2015). Some research has explored the effects of the presence or absence of accuracy information on decision-making quality. Researchers have found that showing subjects a bot's accuracy information led to human performance improving over 20%, compared to showing only the clues used by the bot to make assessments (Lai and Tan 2019). There are also discussions about the differences in observed accuracy and claimed accuracy which have concluded that observed accuracy affects individuals' attitudes about bots' accuracy more than claimed accuracy. For example, algorithms with slower response time improve users' assessments of the algorithm's accuracy, given the same level of

claimed accuracy (Park et al. 2019). In a series of lab experiments, Yin et al. (2019) randomly assigned subjects to one of five accuracy levels: none (the baseline), 60%, 70%, 90%, or 95%, and found that individuals' trust in the model was significantly affected by its observed accuracy, regardless of its stated accuracy.

However, the current body of work on the behavioral implications of communicating accuracy information to users has focused on a direct communication approach, which treats users as passive information receivers. As discussed earlier, users can actively seek out information and generate explanations about a bot's actions, potentially leading to better decision-making than those who receive information passively. Therefore, I investigate how individuals who actively discover a bot's imperfections may respond differently than those who are passive information receivers when the bot's accuracy is low.

In a classification problem such as content moderation, the algorithm predicts whether a comment is appropriate or not. For each comment to be processed, there are two states, ground truth and prediction result. For example, the ground truth of a comment violating comment rules is "inappropriate"; the prediction result from the algorithm, however, may either be a correct prediction ("inappropriate", true positive), or a miss ("appropriate", false negative). Similarly, if the ground truth of a comment not violating any comment rules is "appropriate", the algorithm may either correctly predict it as "appropriate" (true negative), or wrongly treat the comment by predicting it as "inappropriate" (false positive). In the context of content moderation, the accuracy is the proportion of correct predictions (true positive + true negative) out of all processed comments.

Aligning with what academic research has shown about the trust threshold level, Google Perspective, which is deployed in popular news platforms like The New York Times, Reddit, and

The Wall Street Journal, uses a default cutoff of 70% to classify whether a comment aligns with commenting rules (Marvin, 2019). For algorithms operating at a "medium" level (60%, in our case), I propose that users who go through the self-generated explanation will still perceive greater responsibility to monitor the algorithm, and thus detect more errors than the corresponding directly informed groups. However, empirically, decision makers may give up working with the algorithm when the accuracy declines below a certain threshold (for example, chance level: 50%), because they may perceive that the monitoring responsibility is too heavy and be averse to an algorithm of such low accuracy. Instead, they may rely on themselves to make decisions, and the "suggestions" from the algorithm may even backfire and cloud their judgement. I propose that prompting users to provide a self-generated explanation will compensate for the threshold of users' willingness to work with a bot. While the threshold for users being willing to work with a bot when directly informed about its accuracy is around 70%, I believe that active discovery through self-generated explanation will enable users to work with a bot whose accuracy level is below 70%. Thus, I posit:

H3(a): When using an imperfect bot for decision-making processes, users prompted to explain the bot's actions will have higher decision quality, even if the bots' accuracy is low, as long as the bots' accuracy exceeds a certain threshold.

Prior studies have also shown that algorithms' accuracy can also influence individuals' attitudes towards the institutions and authorities that deploy the algorithms (Margetts and Dorobantu 2019). Research has shown that when algorithms are perceived to be accurate and unbiased, people are more likely to trust the institutions and authorities that use them. However, if algorithms are perceived to be inaccurate or biased, this can erode trust in these institutions and authorities (Green 2022). I posit that a bot's perceived accuracy will affect user engagement in an online community, since the online community has the authority to deploy the bot. Thus, I posit:

H3(b): When using an imperfect bot for decision-making processes, users prompted to explain a bot's actions will have less engagement if the bot's accuracy is low. However, this effect will be mitigated if users have actively discovered the bot's imperfections through self-generated explanations, instead of being directly informed.

1.3.3.2 The Algorithm Valence Effect

Algorithm valence refers to the positive or negative emotions that people associate with algorithms. Affective valence can be attributed to the value judgment of an event, such that positive (negative) emotions lead to the perception that the event is pleasant (unpleasant) (Sarker et al. 2005). Prior research has widely discussed individuals' different reactions to stimuli with different valences, and shown that individuals react positively to a pleasant feedback environment (Coursaris et al. 2018, Marvin and Shohamy 2016, Sarker et al. 2005, Willemsen et al. 2011, Yoshida and Yonezawa 2018).

However, fewer studies have tested the behavioral implications of algorithm valence. One exception is a study that tested whether bots should express positive emotions in the context of customer service (Han et al. 2022). However, their interest was to compare how positive emotions expressed by a human versus a bot would be perceived differently by customers. To the best of the authors' knowledge, there is limited research directly comparing a bot developed through a similar approach but with different valence. One possible reason is that existing bot-related studies are within the context of customer service or education, in which it would be inappropriate to design a bot that expressed negative emotions (Feil-Seifer Maja and Matari´c 2011).

Within the context of online content moderation, depending on how platforms deploy the bot, individuals could either reward those they know have been kind to others (i.e., post insightful comments) or punish those they know have been unkind to others (i.e., post inappropriate

comments). When the bot's function is to mark or remove comments violating the discussion rules, its valence is negative ("negative bot"). Users are aware of the comment policy such as "Abusive, defamatory, offensive or disparaging comments on the basis of disability, ethnicity, gender, or otherwise....attacks or threaten another person, promote violence, wish for harm to befall another person...Stalking or harassing another person and any form of discouraging participation by others...Misrepresentation of comments, user profiles, posting advertisements, and all forms of spamming are forbidden." A negative orientation can help users realize their mistakes and avoid them in the future, which improves their task performance (Hattie and Timperley 2007, Kluger and DeNisi 1996). In fact, prior studies have shown that negative feedback may improve performance in tasks such as editing user-generated content (Halfaker et al. 2011, Moon and Sproull 2008, Zhu et al. 2013).

When a bot's function is to highlight and recommend insightful contributions to a community, its valence is positive ("positive bot"). Users are made aware of the comment policy with guidelines such as "Comments represent a range of views and are judged the most interesting or thoughtful. In some cases, comments may be highlighted to showcase commentary from a particular region, or readers with first-hand knowledge of an issue." In the positive bot scenario, users who ignore problems and only praise positive attributes of messages usually fail to censor some of those errors, diminishing the forum's moderation quality. When a positive bot suggests recommending a comment, users may approve it even if they do not fully agree with the recommendation. In other words, if platforms tune the focus of moderation from removing inappropriate comments to identifying insightful content, users will be more likely to agree with a bot. With a negative bot, users will be more cautious and select more correct choices when they

need to confirm a removal decision, compared to confirming a commendation decision from a positive bot.

A bot's positive valence creates a pleasant feedback environment for users, which influences their affective commitment and engagement behaviors (Bateman et al. 2011, Norris-Watts and Levy 2004). Such a favorable feedback environment, where praise and recognition of user-generated comments are prevalent, will lead to more organizational citizenship behaviors among organization members (Becker and Klimoski 1989, Rosen et al. 2006). A positive bot deployed in an online community creates a favorable feedback environment, where praise and recognition of user-generated comments are of high frequency. In contrast, a negative bot's processing orientation is less favorable, because the bot emphasizes punishment, and does not provide compliments for exemplary helpful comments. In the context of online communities, users in a positive and favorable feedback environment will have more enjoyment, joyfulness, and affective commitment, and thus engage more with the community, such as giving more thumbs up, posting more positive comments and replies, and voluntarily helping the bot to manage content. Thus, my final set of hypotheses is:

H4: All else equal, a positive bot increases a user's engagement, compared to a negative bot.

In addition, when confronted with two types of bot errors arising from positive or negative bots, users may react differently. If a negative bot makes a false negative error (i.e., misses an inappropriate comment and marks it as ok), individuals are motivated to protect the community, correct the bot's error, and override the bots' assessment; if the negative bot commits a false positive error (i.e., punishes an innocent comment), individuals are motivated to speak up for the "victim" and report the bot's incorrect evaluation. If a positive bot commits a false negative error

(i.e., misses an insightful comment), individuals are motivated to acknowledge the contribution and recommend the comment to the bot. However, if the negative bot commits a false positive error (i.e., marks an ordinary comment as insightful), individuals may simply remain silent because individuals are reluctant to recant existing compliments. This is because individuals expect others to treat themselves in a similar way that they treat others (Nowak and Sigmund 2005). The generosity to echo the bots' reorganization of a comment, even if the comment is not insightful at all, aligns with social interaction norms (Hawkins et al. 2019). In short, the decreased decision making is mainly due to the decrease in positive errors caught by positive bots. It remains an open question how individuals' moderation behaviors will vary if working with bots with either positive or negative valence.

1.3.4 The Underlying Mechanism: Complacency Potential

Complacency, or sub-optimal monitoring of automation performance, has been cited as a contributing factor in numerous major transportation and medical incidents involving humans and algorithms (Bahner et al. 2008, Chan Fung 2021). Complacency potential refers to the degree to which automation can lead to a decrease in vigilance and attention to task performance. Complacency has been conceptualized and measured through two dimensions in prior research (Merritt et al. 2019, Parasuraman and Manzey 2010): (a) Alleviating Workload: the attitude about using automation to ease workloads, and (b) Monitoring: the degree of lack of attention to monitoring automated tasks.

Research has shown that automation itself can lead to complacency because of the blurred lines of responsibility. For example, Parasuraman and Riley (Parasuraman and Riley 1997) found

that pilots using an automated flight control system had a higher risk of crashes due to complacency.

One important factor that can influence the complacency potential is the degree of task automation. Research has shown that the higher the degree of automation, the higher the potential for complacency. For example, Endsley et al. (2015) found that operators of highly automated systems in a nuclear power plant were more likely to exhibit complacency than those operating less automated systems.

Moreover, the complexity of the task and the user's level of expertise can also influence the complacency potential. Research has shown that less experienced operators are more likely to exhibit complacency when using automated systems. For example, Raguphthi and Hass (2011) found that novice surgeons using a robotic surgery system were more likely to exhibit complacency than expert surgeons.

Additionally, the type of automation and feedback provided can also influence the potential for complacency. Research has shown that automation that provides little feedback or that requires little human input can lead to higher complacency. For example, Sebok and Wickens (2017) found that operators using a fully automated system with little feedback were more likely to exhibit complacency than those using a system that required more human input.

Prior research has shown that complacency is a heuristic replacement for monitoring automation performance (Singh et al. 1993). It is well documented that humans tend to choose the road of least cognitive effort in decision making, and instead of basing complex decisions on a comprehensive analysis of all available information, users often use simple heuristics and decision rules (Betsch et al. 2004, Tversky and Kahneman 1974). Bots' recommendations may serve as a strong decision-making heuristic for human users and may substitute for more effortful

information analysis and evaluation processes (Mosier and Skitka 1996). In addition, sharing decision-making tasks with a bot may lead to the same psychological effects that occur when humans share tasks with other humans. For example, "social loafing" or the tendency to reduce one's own effort when working within a group may occur when users work collaboratively with a bot (Canning and Harackiewicz 2015, Hitron et al. 2019). To the extent that human users perceive bots as another team member (Brown et al. 2010, Rai et al. 2019), they may see themselves as less responsible for the outcomes, and as a consequence, users will reduce their own efforts in analyzing all available information and become complacent in monitoring the bots' actions.

I propose that generating an explanation may reduce complacency by triggering a higher perceived sense of responsibility to monitor an algorithm, and thus increase the error detection rate. First, through a self-generated explanation exercise, decision makers expect an algorithm to fail in the tasks that challenge it. Individuals who are actively involved may perceive an increased responsibility to monitor an algorithms' suggestions and conclusions and anticipate its errors. Such perceived responsibility will motivate users to process all the contextual information and algorithms' suggestions more thoroughly. Thus, when algorithms actually commit errors, they will be able to detect them and override systems' decisions. Furthermore, those users would have increased awareness that it is their responsibility to handle certain challenging cases for the algorithm. The complacency and social loafing happen when individuals perceive they can rely on the algorithm. However, when they anticipate an algorithm will commit errors under certain circumstances, they will perceive more responsibility to monitor the algorithm and intervene if necessary. Thus, I posit:

H5: The relationship between self-generated explanation and decision-making quality is mediated by reduced complacency potential.

1.4 Experiment Design: Overview

Based on the theoretical framework (Figure 1) and corresponding hypotheses, I conducted five experiments to examine the collaborative workflow of human-AI hybrid design. In Study 1, I examined whether prompting users to provide self-generated explanation about bots' actions will achieve higher decision-making quality compared to users getting exposed to bots' imperfections through direct communications (H1); In Study 2, I replicated and extended Experiment 1 by exploring whether the effects of self-generated explanation endure (H2); In Study 3, I explored how an algorithm's accuracy changes the relationship between self-generated explanation and decision quality (H3); In Study 4, I explored how users' engagement changes due to algorithm's valence (H4); In Study 5, I replicated and extended Experiment 1 by exploring reducing complacency potential as the underlying mechanism of the impacts of self-generated explanation and decision making quality (H5). For all five studies, I collected data in single, complete batches and did not conduct any analyses until all the data for a given experiment were collected. My total final sample size across five experiments was 1650 participants (see Table 2). Participants who were undergraduate students were recruited from a large state university in the United States; they were registered subjects in the University's behavioral lab and were awarded course credit for their participation. MTurk subjects refers to adult subjects recruited through Amazon Mechanical Turk.

Table 2 Data Collection Waves

Data collection wave	Subject	Number of subjects
1	MTurk	175
2	Student	278
3	MTurk	500
4	Student	310
5	Student	221
6	Student	166

1.4.1 Experiment Setup

1.4.1.1 Collaborative Human-AI Task: Content Moderation

The collaborative human-AI task in our experiments was to moderate comments with assistance from an AI bot in a simulated news discussion forum. For comments that the bot assessed as inappropriate based on forum discussion rules, a message was displayed next to the comment, reading "Bot assessment: Not Appropriate". Subjects were able to click "agree" or "disagree" buttons to provide feedback about the bot's assessments, and report any comments they believed the bot had missed. There was one mandatory and one voluntary moderation task for subjects to complete in the forum. To complete the mandatory task, subjects performed the following steps: log into the discussion forum, read the selected news article, learn about the community's rules on commenting, post a comment about the article, receive training on the commenting rules, learn how the moderation AI bot works, and finally, rate five peer-generated comments in sequence.

After completing the mandatory moderation task, subjects entered the full discussion forum, which had been populated with new comments (twelve in Experiments 1 and 3, and twenty-four in Experiment 2). At this stage, participants were able to interact with the bot as well as other users' comments on a voluntary basis. While engaging with the discussion forum, subjects were able to click "agree" or "disagree" buttons to provide feedback about the bot's assessments, and report any other comments to the forum that they believed the bot had missed. In addition, subjects were free to engage with the forum as much as they liked, including, but not limited to, leaving comments, replying to existing comments, and up-voting and down-voting comments.

To observe the subjects' content moderation performance without the bot's assistance, we also deployed a *NoBot* condition as a baseline. In this baseline condition, for the mandatory

moderation task, users were requested to make a binary moderation decision about whether or not to publish a comment. For the voluntary task in the baseline condition, subjects could freely use the report button to mark any user comments that they deemed inappropriate.

1.4.1.2 Experiment Materials and Procedural

The news article and the related comments used in the experiments' discussion forms are from Yahoo News. I purposely chose a controversial topic in the politics section: recreational marijuana legalization, because politics is known to attract participation, and sometimes attract very divisive user comments (Coe et al. 2014). In order to reduce subjectivity, I chose a fact reporting style to summarize the new laws on marijuana in Michigan, Missouri and Utah (Keenan 2017). I used the Yahoo News API to collect all the comments of the article and selected comments to be used in the experiment. Among those selected comments, the goal was to have a balanced mixture of the quality (insightful, abusive, and normal comments) and attitude towards legalization of recreational marijuana use (for and against). Three authors classified the attitude of each comment to be "for" or "against" recreational marijuana legalization independently. I then used the Google Perspective API to assess the toxic quality of each comment. Google Perspective uses machine learning models to score the toxicity level of a comment and identifies whether a comment could be perceived as toxic to a discussion. The Google Perspective API is currently used by online communities including Wikipedia, The New York Times, The Economist, and the Guardian (Blue 2017).

Comment boards of online news media generally show six to ten comments at one time to the users. Considering this observation from the real world, and the information load I needed to manipulate in my experiment, I eventually displayed twelve comments on one screen. In Study 5, I extended the number of comments to twenty-four. The selected comments had equal amounts of

agreement and disagreement about recreational marijuana legalization (six comments each, for Studies 1-4; twelve comments each, for Study 5), and abusive, insightful, and normal comments (four each, for Studies 1-4; eight comments each, for Study 5). In order to make sure users saw all comments and moderation reports from the bot (if applicable), I adopted a jump-down design: when users posted their comments in the comment box, the web page automatically led the user to the bottom of the page to see their own posted comments. The comments were consistent across experimental waves, but the order of comments for each subject is random.

In order to simulate actual practices used to regulate comment boards, I adopted a few designs that have been shown to be effective for managing online discourse in my experiment. For example, participants were requested to give themselves a screen name (a username to be displayed on the site). When they were redirected to the discussion page (before seeing the news article and the comment board), they needed to first log in with their screen name and a series of numbers provided by the researchers as a log-in password. Each participant's screen name and simulated avatar or photo were shown on the web page when they posted any comments or replies. For the comments in the discussion forum, I kept the real usernames that I had collected from Yahoo.com. I showed a key instruction on the comment board ("Please be polite and read our guidelines") ahead of the comments section. Users were also able to re-check the comment rules by hovering over a question mark icon on each assessed message from the bot.

To avoid subject-expectancy effect, participants were not told of the experiment's purpose in advance and were only told that the study was aimed at collecting public opinions about marijuana legalization. After the experiment, participants were fully debriefed.

Table 3 Experiment conditions in all five studies

Study 1	Study 2	Study 3	Study 4	Study 5
NoBot	NoBot	NoBot	PositiveBot	NoBot
InformBot	InformBot	InformStrictBot	NegativeBot	DisclaimerBot
ExplanationBot	ExplanationBot	InformLenientBot		AlertBot
	DelayedExplanationBot	ExplanationStrictBot		InformBot
		ExplanationLenientBot		ExplanationBot
		EnhancedExplanationStrictBot		
		EnhancedExplanationLenientBot		

Table 3 lists all the experimental conditions across the five experiments, and Table 4 provides a summary of the experimental task workflow for all the treatment groups. Among all the nine groups, I labeled the following five groups as "explanation" groups, because they have an explanation prompt component in their workflow: (1) ExplanationBot, (2) ExplanationStrictBot, (3) ExplanationLenientBot, (4) EnhancedExplanationStrictBot, (5) EnhancedExplanationLenientBot. Users in all the experiment groups were informed about the discussion forum's commenting rules. When the bot was present, they were also informed about the role of the bot as a comment moderator and presented with four example comments with bots' assessments. Next, I presented the same details about the bot's classification approach and highlighted the bots' errors in the examples that were already seen by the subjects. Only subjects in the explanation groups were prompted to explain how they thought the bot operated and why they thought the bot came to the particular moderation conclusions that were shown in the examples.

Table 4 Experimental Tasks Workflow

Experiment workflow items	*NegativeBot*, *PositiveBot*, *ExplanationBot*, *ExplanationStrictBot*, *ExplanationLenientBot* condition	*Enhanced ExplanationStrictBot* and *Enhanced ExplanationLenientBot* condition	*InformBot*, *InformStrictBot*, and *InformLenientBot* condition	*NoBot* condition	*Delay Explanation Bot* condition
Log into discussion forum	1	1	1	1	1
Read the article	2	2	2	2	2
Receive training on commenting rules	3	3	3	3	3
Draft and save a comment	4	4	4	4	4
Learn how the bot works and perform a pause and generated a self-explanation about bots' actions *(i.e., "Why are comments 1 and 3 assessed as not appropriate whereas comments 2 and 4 are appropriate? ")*	5	5	NA	NA	5
Another pause and self-explanation task *(i.e., "Did the bot make any mistakes in assessing comments 1-4?")*	NA	6	NA	NA	NA
Rate 5 user-generated comments and register agreement or disagreement with bot (for *NoBot* condition, decide to publish the comment or not)	6	7	5	5	6†
Re-read stored comment draft and optionally revise the comment before final posting to the forum	7	8	6	6	7
Enter and engage with the discussion forum and perform voluntary actions. The forum was seeded with 12 new comments, and participants freely engaged with the forum	8	9	7	7	8†
Debrief and end of experiment	9	10	8	8	9

1.4.2 Measurement

One group of desired outcomes is moderation decision quality. Three dependent variables of interest were used to measure the decision-making quality of the content moderation tasks. I measured decision-making quality as the proportion of correct assessments in the binary moderation decisions made by a user (whether a comment is in compliance with the forum's rules or not). The first dependent variable, Detection of bot's error rate, refers to the error detection rate of the bots' errors, and the second dependent variable, Mandatory task decision quality, refers to the accuracy of a user's moderation decisions in the mandatory rating task.

The third variable, voluntary task decision quality, refers to the accuracy of a user's moderation decisions on the forum, where users may interact with the bot on the discussion forum. In Studies 1-4, there were six bots' errors to be corrected (three inappropriate comments being missed by the bot and three appropriate comments being wrongly marked). Therefore, one way to measure voluntary moderation quality is to divide the number of correct decisions by the total number of errors (i.e., six). I refer this measurement as *forumAccuracy*. Such measurement, however, only captures user' behavior with the six comments that were designed to have bot errors, and misses users' interaction with the rest. In addition, because users interacted with the bot in a voluntary manner, the total number of voluntary moderation decisions varies across subjects. Thus, an alternative measurement is to divide the number of correct decisions by the total number of users' actions, which I refer as f*orumAccuracySelf*. For example, if a user chose to make 5 moderation decisions (3 with the target comments), and 2 of these decisions are correct (1 with the target comment), then the *forumAccuracy* is 0.25 (1 divided by 4), and *forumAccuracySelf* is 0.4 (2 divided by 5). Please note that In Study 5, there were 24 comments in the forum and 12 bots'

errors to be corrected (six inappropriate comments were missed by the bot and six appropriate comments were wrongly marked).

I operationalized user engagement in two ways, including total interaction count, changes in subjects' comment sentiment, and the count of voluntary moderation. For the total interaction count, I calculated the engagement by counting all users' actions in the forum except those related to moderation, including thumbs-up or thumbs-down for a comment or reply and relying on a comment or reply. Alternative methods to calculate users' engagement level are to assign weights to different actions, instead of treating all the actions as equal and summing them up (Cheng et al. 2017). A main motivation to assign weights to different actions is because expressing-related actions (such as replying or commenting) are perceived as more engaging than those who express themselves in a salient mode (such as clicking thumbs up or down) or simply browsing. However, theoretically, there is no clear indication about how the weights should be decided. In addition, due to the limited number of comments available in the forum (12 for Studies 1-4 and 24 for Study 5), assigning weights may easily lead to a high disparity in the measured engagement and biased results. Thus, in my current dissertation, I chose to use the total number of actions to represent engagement level. Please note that clicking the "report" button to report a missing inappropriate message to the bot or forum was excluded because that was captured in the voluntary moderation accuracy.

To observe the effect of self-explanation on subjects' comment sentiment change, I compared the sentiment of subjects' revised comments (after the treatment) to their original comments (before the treatment). For comment sentiment analysis, I used the Python package AFINN (Nielsen 2011) to calculate sentiment scores for comments. The AFINN, however, has several limitations, such as a limited vocabulary of words that are assigned scores based on their

sentiment. In addition, the AFINN does not take into account the context in which a word is used (Srivastava et al. 2022). For example, the word "killer" may have a negative connotation when used to describe a person, but may have a positive connotation when used to describe an effective product. To provide a robustness check for the results, I also calculated the sentiment scores with the NLTK package.

For deriving the total quantity of voluntary moderation decisions of a user, I counted the number of voluntary clicks on the report, agree with bot, and disagree with bot buttons that were made while the user freely engaged with the forum.

In Study 5, I also deployed an embedded mouse movement tracking service on the webpage to track subjects' mouse trajectories through FullStory (https://www.fullstory.com/). I used a host of pre-treatment variables as control measures, including users' general news reading habits, abuse reporting habits, attitudes toward recreational marijuana legalization, and demographic information (age, gender, race, and education). I collected this information from the participants using a survey that was administered before the start of the experiment.

Table 5 List of Variables

Variable	Description
Panel A. Pretreatment variables	
Gender	Subjects' gender
Age	Subjects' ages in years
Race	Subject's race
Education	Subject's education background
Income	Subject's income level
Marital	Subject's marital status
Know_LM	Subject's knowledge level on the legalization of marijuana
Vote_LM	Subjects' attitude on the legalization of marijuana (for or against)
Read news	Subjects' online news reading frequency
Read Comments	Frequency of reading comment boards when reading online news
Comment	Frequency of posting comments when reading online news
Reply	Frequency of replying to others' comments when reading online news
Vote	Frequency of clicking like or dislike for comments when reading online news
Report	Frequency of reporting inappropriate comments when reading online news
Version	Wave of data collection
Panel B. Post-treatment variables	
Cognitive effort	Seconds spent on the moderation task
Mandatory moderationdecision quality	Accuracy of the moderation task
Bot error detection rate	The percentage of correct disagreements with bot's wrong suggestions over the total number of bot suggestions
Voluntary moderation decision quality	Accuracy of voluntarily moderated comments during interactions with the discussion forum
Total interaction count	Count of actions in the forum (moderation-related excluded)
Comment sentiment change	The sentiment changes of user-generated original comments and revised comments
Count of voluntary moderation	Count of the number of voluntary moderation actions on the forum

2.0 Study 1: Performance Effects of Directly Communicated and Self-Generated Explanations about Algorithmic Errors

The purpose of Study 1 is to explore the moderation decision quality differences for users who were directly told about bots' imperfections or actively discovered it through generating a self-generated explanation (H1).

2.1 Method

I aim to compare the impact of user-generated explanations of bot actions (*ExplanationBot*) on decision-making quality with a commonly used approach that organizations use to communicate bot errors directly to users: Examples of bot errors with direct pointing (*InformBot*). To observe users' moderation decision quality without any bot assistance, I also included a baseline group: 3) *NoBot*. All participants were provided with the rules for commenting on the forum. Those who were not assisted by the bot (*NoBot condition*) were simply notified that all comments on the forum would be moderated, and they were requested to rate five comments as part of a moderation task before entering the full discussion forum. In contrast, those in the bot-assisted conditions (*ExplanationBot, InformBot*) were notified that the forum uses a bot for assisting moderation of content.

The subjects in *InformBot* and *ExplanationBot* learned about how the bot worked and read four specific comment examples that were categorized by the bot as appropriate or inappropriate. In the experiment, a key aspect is to be aware of and be attentive to the actions of the imperfect

algorithm. As noted earlier, I used the self-generated explanation technique to trigger a subject to anticipate algorithmic errors (Williams and Lombrozo 2010). After seeing four example comments, subjects in the *ExplanationBot* group were first asked to generate an explanation on how they thought the bot operated and why they thought the bot came to the particular moderation conclusions that were shown in the examples. After participants submitted their explanations, I illustrated that the bot is imperfect and can make mistakes. As shown in Figure 2, I categorized one example comment in each cell of the classification matrix: (1) a true positive, (2) a true negative, (3) a false positive (type I error), and (4) a false negative (type II error). In contrast, in the directly display condition (*InformBot*), after seeing the four examples, users were directly displayed the imperfections of the bot that were revealed through examples. For both groups, the same amount of information, including details about the bot's classification outcomes and the explanations about bots' errors that were presented. The only difference between the *InformBot* and *ExplanationBot* groups is the latter was prompted to provide an explanation about why the bot assessed comment 2 and comment 4 as appropriate, and comment 1 and 3 as inappropriate. While they were providing the explanations, I expect them to be aware that the bot made mistakes in classifying comment 3 and 4. Later, when they read the information in Figure 2, their suspicions would be confirmed. Because users in the explanation treatment condition (*ExplanationBot*) paused and spent more effort explaining and learning about the classification approach of the bot, they tend to be more aware of and attentive to the actions of the bot, and have a higher error anticipation, relative to the *InformBot* group.

		Assessment by human moderators	
		Inappropriate	Appropriate
Predicted by Bot	Inappropriate	**Comment -1:** *Climate change is happening and it's not changing in our favor. If you think differently you're an idiot.* Bot is correct. An inappropriate comment is caught by bot (same with the human moderator)	**Comment -3:** *Some are just poorly educated, ultimately not their fault for being uninformed and ignorant. I blame the American educational system.* Bot is wrong. Comment classified as inappropriate by the bot when it is appropriate
	Appropriate	**Comment -4:** *You either trust in God or think you are smarter than him as you believe in this crooked science where there is no consensus.* Bot is wrong. Comment classified as appropriate by the bot when in reality the comment is NOT	**Comment -2:** *Clearly man made, but unsure of its extent and whether anything substantial can be done about it* Bot is correct. An appropriate comment is identified by bot (same with the human moderator)

Figure 2 Example comments to Highlight Bots' Errors to Users

As noted earlier, I used the self-explanation technique to trigger a subject to be more aware of the need to monitor the system (Roscoe and Chi 2008, Williams and Lombrozo 2010). In my experiment, a key aspect is to be aware of and be attentive to the actions of the imperfect bot. In the Inform condition (*InformBot*), users were informed about the existence of the bot, and the imperfections of the bot were revealed through examples, but the users were not asked to pause and explain the inner workings of the bot. In contrast, subjects in the *ExplanationBot* group were first asked to explain how they thought the bot operated and why they thought the bot came to the particular moderation conclusions that were shown in the examples. After participants submitted their explanations, I presented the same details about the bot's classification outcomes and the explanations about bots' errors that were seen by the subjects in the *InformBot* condition. Because users in the explanation treatment condition (*ExplanationBot*) paused and spent more effort explaining and learning about the classification approach of the bot, they tend to be more aware of and attentive to the actions of the bot, relative to the *InformBot* group.

2.2 Results

2.2.1 Manipulation Checks

To check whether the explanation prompts indeed triggered users' active involvement, I compared the total time spent on the content moderation tasks across the experimental groups. As shown in Figure 3, *NoBot* subjects spent, on average, about 164 seconds completing the moderation task. Similarly, subjects with the bot's assistance but without self-explanation manipulations (*InformBot*) spent about 178 seconds on the moderation task. Subjects with self-explanation manipulations (*ExplanationBot*), however, spent much more time assessing the moderation tasks (about 226 seconds), compared to all the other groups. Regression results further corroborated that the level of cognitive effort expended by subjects in the *ExplanationBot* condition is significantly higher than all the other groups.

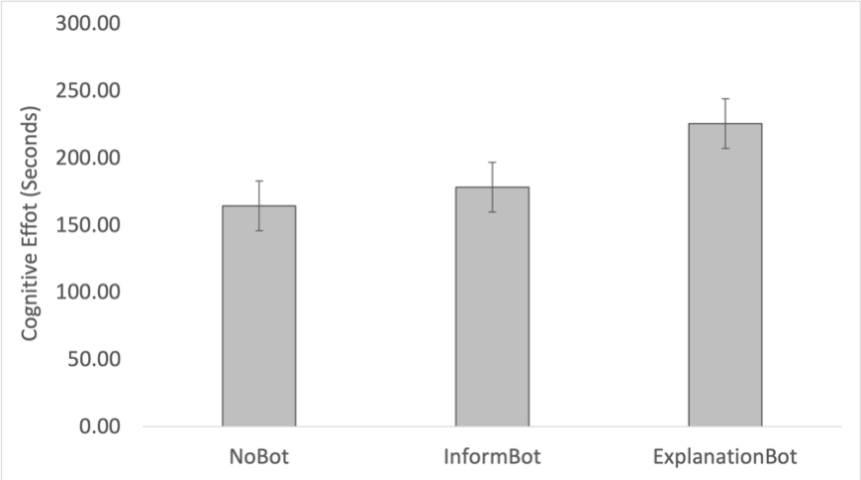


Figure 3 Bot, Self-Generated Explanation, and Cognitive Effort (Study 1)

2.2.2 Self-Generated Explanation Prompt Improve Decision Making Quality

I first compared the cell means of each group to provide model-free evidence for the effect of bot and explanation prompt on subjects' decision-making quality. From Figure 4, I can see that subjects who encountered the bot and were prompted to provide explanation (*ExplanationBot*) had the highest level of decision quality (72.1% accuracy), followed by the *InformBot* group (59.5% accuracy). The subjects in the *NoBot* group exhibited the lowest level of decision quality (13.3% accuracy). Table 6 provides the corresponding regression results for testing H1(a), H2(a), and H3. As shown in Model 1 of Table 6, *NoBot* subjects had a 13.3% accuracy rate. With the bot's assistance, the accuracy of both the *ExplanationBot* and the *InformBot* groups increased significantly, relative to the baseline *NoBot* group. The coefficients of *ExplanationBot* and *InformBot* are significant in the models and the results of the tests comparing the coefficients also reveal statistically significant differences. As presented in Model 2 of Table 6, the results are consistent when the models include a host of pretreatment variables as covariates. Thus, H1(a), predicting an increase in decision making quality when highly accurate but imperfect bots are used as decision aids, is supported in the data. The results also indicate that the pause-and-explanation practice resulted in about 21.2% higher decision quality, relative to those who did not receive explanation prompt, which lends strong support to H1(a).

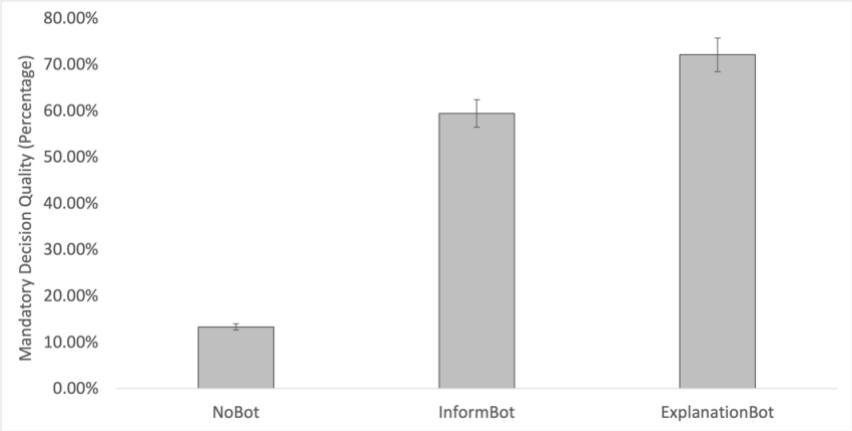


Figure 4 Bot and Decision Quality (Study 1)

Note. Error bars indicate 95% confidence interval of the mean

Table 6 Bot and Decision Quality (Study 1)

Variables	Model 1	[95% Conf.	Interva]	Model 2	[95% Conf.	Interval]
InformBot	0.461***	0.391	0.532	0.46***	0.397	0.522
ExplanationBot	0.588***	0.515	0.661	0.589***	0.521	0.658
Constant[#]	0.133	0.081	0.185	0.473	-0.309	1.256
Covariates[+]	No			Yes		
Coefficient comparison	InformBot - ExplanationBot = 0			$\chi^2(1) = 14.84$ Prob > F = 0.000		

2.2.3 Heterogeneous Treatment Effects on Decision Making Quality

As results in Table 6 demonstrate, the *ExplanationBot* treatment had the highest positive impact on decision quality. To explore the heterogeneity of this treatment effect on subjects, I analyzed nonparametric causal trees, which are regression trees used for the prediction of treatment effects. Causal trees partition the covariate space into a decision tree that minimizes prediction error while estimating constant treatment effects within each leaf of the tree (Athey et al. 2017). I used the causal tree method to evaluate how subjects respond differently to self-explanation prompts. Specifically, I used the *causalTree* package in R for analysis and used tenfold cross-validation to avoid overfitting (Athey et al. 2017). As shown in Figure 5, I find that the impact of self-explanation prompts depends on age, prior forum interaction habits, topic knowledge, and

topic related attitude. Specifically, subjects who are younger than 20 (34% of the sample) are less likely to be influenced by the self-explanation treatment (their accuracy in classifying messages was 4.6%). Subjects who are older than 20, with a higher level of prior knowledge on the topic, and with less prior experience with discussion forums benefitted substantially more from the *ExplanationBot* treatment (their accuracy in classifying messages was 71%).

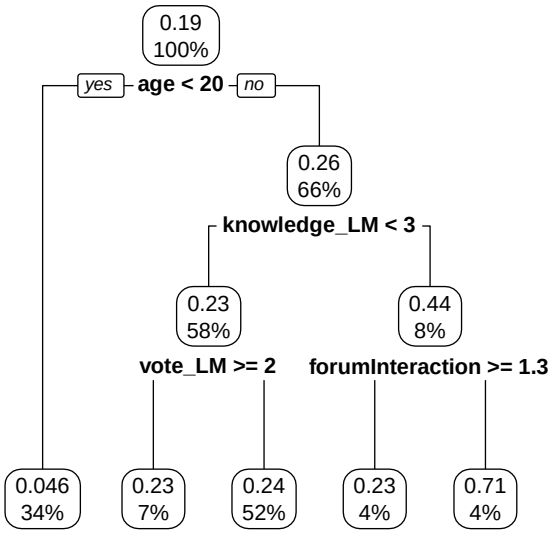


Figure 5 Heterogeneity of the Self-Explanation On Decision Quality (Study 1)

Note. The Root Mean Square Error (RMSE) of the prediction using the causal tree is 0.49

2.2.4 Self-Generated Explanation Prompt and Detecting Bot Errors

I proceeded to explore the underlying mechanism behind the improvement in the decision quality of subjects in the *ExplanationBot* condition. In my hypothesis development, I posited that users who provide explanation would scrutinize the suggestions of imperfect bots and would be able to better detect errors made by the bots, relative to the users in the *InformBot* condition. Figure 6 presents evidence for this underlying mechanism and indicates that subjects with self-explanation prompts were able to detect and rectify 62.1% of the imperfect bot's errors, but users

in the *InformBot* condition only identified 39.3% of the bot's errors. Regression results corroborated these significant differences, and I find strong support for my postulation that self-explanation prompts help users with better error detection and correction.

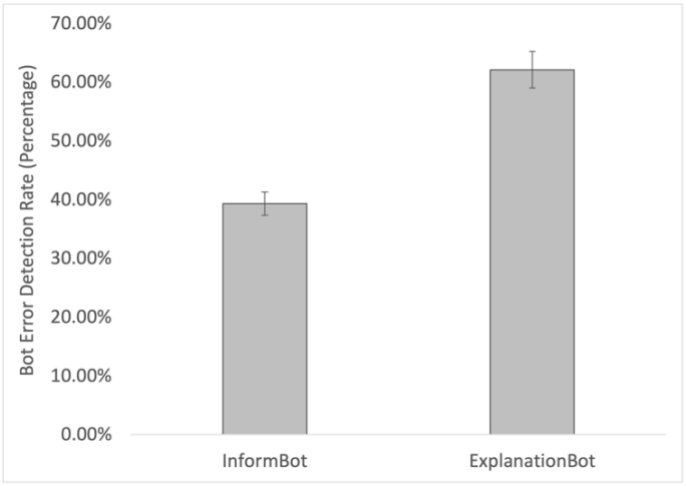


Figure 6 Detect Bot's Error Rate (Study 1)

Note. Error bars indicate 95% confidence interval of the mean

2.2.5 The Effects of Bot and Self-Generated Explanation Prompt on User Engagement

One of my goals is to examine how subjects voluntarily contribute to the discussion forum and help the community to manage comments in the discussion forum. H1(b) predicts that engagement will increase with the use of a moderation bot and with self-explanation prompts.

I first examined whether the presence of the bot and explanation prompt could influence total interaction counts, trigger subjects to revisit their comments, and to engage in more voluntary moderation decisions. As the results reported in panel A and B of Table 7 show, I did not find evidence supporting my expectations for overall forum interactions and comment sentiment change. There are no significant differences between the subjects in the *NoBot* and the other three

groups with respect to comment sentiment change. Upon reflection, I think a possible reason for this result could be that the bot was not designed to specifically encourage people to contribute or engage more, and the design focused on nudging subjects to locate inappropriate comments. The results in the Panel C of Table 7 reveal that the presence of the bot in both the *ExplanationBot* and *InformBot* groups decreased users' intention to help make voluntary moderation decisions while voluntarily engaging with the discussion forum.

Table 7 Effect of Self-explanation on Engagement (Study 1)

Panel A: Overall Interactions						
Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	0.959	-1.239	3.157	1.5	-0.761	3.761
ExplanationBot	-1.659	-3.925	0.606	-1.415	-3.730	0.901
Constant[#]	8.508***	6.888	10.128	3.023	-19.938	25.984
Covariates[+]	No			Yes		
Panel B: Comment Sentiment Change						
InformBot	-0.644	-1.862	0.573	-0.441	-1.681	0.799
ExplanationBot	0.525	-0.667	1.718	0.711	-0.559	1.981
Constant[#]	-0.222	-0.849	0.404	-2.168	-14.760	10.424
Covariates[+]	No			Yes		
Panel C: Voluntary Moderation Count						
InformBot	-0.998***	-1.507	-0.489	-1.048***	-1.564	-0.532
ExplanationBot	-1.617***	-2.142	-1.092	-1.679***	-2.207	-1.151
Constant[#]	2.571***	2.196	2.947	-1.671	-6.906	3.565
Covariates[+]	No			Yes		

Note: *N=204; *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of them are statistically significant at p<0.05.*

I also explored the effect of the bot and self-explanation on accuracy of the voluntary moderation behaviors. As shown in Table 8, I found that when a bot was not present, subjects correctly reported, on average, 8.9% of the inappropriate comments. Introducing a bot improved subjects' voluntary moderation accuracy. Specifically, with the aid of the bot, subjects with self-explanation prompts correctly reported 63% of the inappropriate comments, achieving about seven times better performance than that of subjects in the no bot condition. Users who were in the *InformBot* treatment also had better performance than the *NoBot* group, reporting about 47.6% and

49.7% of the inappropriate comments found in the forum, respectively. The differences in voluntary moderation accuracy levels between the *ExplanationBot* and *InformBot* groups are all statistically significant. Overall, the presence of a bot and the self-explanation treatment increased subjects' voluntary moderation accuracy, but it did not significantly improve their overall volume of contributions to the discussion forum. Thus, I did not find sufficient evidence in favor of H1(b).

Table 8 Voluntary Moderation Accuracy (Study 1)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	0.387***	0.320	0.455	0.384***	0.328	0.439
ExplanationBot	0.541***	0.472	0.610	0.542***	0.480	0.605
Constant[#]	0.089***	0.039	0.138	0.564	-0.476	1.604
Covariates[+]	No			Yes		
Coefficient comparison	InformBot - ExplanationBot = 0			$\chi^2(1) =$ 22.12, Prob > chi2 = 0.0000		

Note: *N=204. *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of these covariates are significant at p<0.05.*

3.0 Study 2: Does the Self-Generated Explanation effect endure?

In Study 1, I found that compared to directly inform individuals about bots' imperfections, prompting users to provide a self-generated explanation is beneficial for them to detect more algorithmic errors and increase their moderation decision quality. In Study 1, however, I found such effect only applied to the mandatory moderation task, but not the subsequent voluntary moderation task. There are two potential reasons: First, the active discovery about bot' imperfections only increase users' awareness to monitor the bots' errors if they are required to complete a task. If the task is voluntary, individuals lost willingness working with the bot or may simply rely on the bot to complete the moderation task. Another potential reason is, the reason why users failed to detect errors in the voluntary moderation task is they became tired and their cognitive resources were depleted. To test which explanation is the reason for unsatisfying performance in voluntary moderation task, in Study 2, I replicated and extended Experiment 1 by exploring whether the effect of self-generated explanation endure (H2). In particular, after the bot imperfection training (direct communication or generating self-explanations), I switched the sequence of voluntary moderation task and mandatory moderation task. That is, after the prompt, users first entered the forum and completed the voluntary moderation task, and they were not required to complete the mandatory moderation task until they finished interacting with the forum. Under this setting, if the results showed a performance increase in the voluntary task, but the mandatory task, I could infer that the benefits of self-generated explanation does not endure multiple tasks because users' cognitive resources were depleted.

3.1 Method

In addition to the prior three conditions (*NoBot*, *ExplanationBot*, *InformBot,*), I deployed a fourth experimental condition *DelayExplanationBot*. The only difference between the *ExplanationBot and DelayExplanationBot* groups pertains to the time delay between receiving explanation prompts and the execution of content moderation and forum engagement tasks by the subjects in those groups. Specifically, after learning how the bot works with self-generated explanations, participants in the delayed explanation group (*DelayExplanationBot*) skipped the comment rating task temporarily and proceeded to the discussion forum directly. They completed the comment rating task in the last step, only after finishing their engagement with the discussion forum. Thus, while the subjects in the *ExplanationBot* condition approached the rating task immediately after the self-explanation, users in the *DelayExplanationBot* condition had to endure an intermediate task (engaging with the forum) before they could complete the rating task, which facilitates my assessment of whether the self-explanation could produce positive effects over multiple tasks.

3.2 Results

3.2.1 Descriptive Analysis and Manipulation Check

310 subjects (170 males and 140 females) participated in three experimental conditions focused on bot configuration. Subjects' ages varied from 18 to 29 years, with a mean of 20.1 years old. A majority of the subjects (78.7%) were white. 84.19% of the subjects considered themselves

as having some but not a lot of knowledge on the topic of legalized recreational marijuana. Among all the subjects, 259 supported legalization (83.55%), and 51 subjects held the opposite opinion (14.45%). Regarding their online news reading habits, 79.03% regularly read news online, and 49.68% frequently read comments under news articles, sometimes posting a comment on forums. 80.97% subjects never or seldom post any comments or replies, and only 54.84% of the subjects reported having clicked 'thumbs up' or 'thumbs down' buttons during real-world discussion forum interactions. Notably, a majority of the subjects (78.71%) said they never flagged or reported any inappropriate comments during their real-world forum interactions.

As shown in Table 9, the level of cognitive effort expended by subjects in the *ExplanationBot* condition is significantly higher than all the other groups. The delayed bot group (*DelayExplanationBot*) spent the least amount of time on the moderation task (about 155 seconds), consistent with a drop in the subjects' cognitive resources as they completed the last step of their experiment tasks.

Table 9 Cognitive Effort Spent on the Moderation Task (Study 2)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	13.956	-9.869	37.780	17.871	-9.855	45.597
ExplanationBot	61.207***	36.652	85.762	61.794***	32.480	91.107
DelayExplanationBot	-9.909	-32.086	12.268	-10.313	-31.417	10.792
Constant[#]	164.444***	146.881	182.008	3.187	-438.179	444.553
Covariates[+]	No			Yes		
Coefficient comparison	InformBot - ExplanationBot = 0			$\chi^2(1) =$ 9.52, Prob > chi2 = 0.002		
	InformBot - DelayExplanationBot = 0			$\chi^2(1) =$ 38.53, Prob > chi2 = 0.000		
	ExplanationBot - DelayExplanationBot = 0			$\chi^2(1) =$ 6.44, Prob > chi2 = 0.011		

3.2.2 Effect of Self-Explanation Failed to Endure Multiple Tasks

To test whether the effect of explanation prompt endures multiple tasks, I compared the *DelayExplanationBot* group with the other conditions, and the results show that although the accuracy of these subjects (48.5%) is still higher than the *NoBot* condition (13.3%), it is

substantially lower than those in the *ExplanationBot* and *InformBot* conditions. The tests of coefficient comparisons presented in Table 10 that the 23.6% difference in decision quality between the *DelayExplanationBot* and *ExplanationBot*, and the 11% accuracy difference between the *DelayExplanationBot* and *InformBot* conditions are significant at the $p<0.05$ level. These results indicate that the effect of self-generated explanation deteriorates over multiple tasks, and thus, I find evidence supporting H2. As a robustness check, as shown in

Table 11, I used an alternative measure of decision-making quality, the "F1" score, another commonly used metric to evaluate the performance of a binary classification model. Compared to Accuracy, which only reflects correctly classified instances out of all instances, the F1 score takes into account both precision and recall, and is especially useful when the class distribution is imbalanced. Similar to accuracy, the F1 score also ranges from 0 to 1, with 1 being the best score possible. I found similar evidence supporting H1(a), and H2.

Table 10 DelayExplanationBot and Decision Quality – Accuracy (Study 2)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	0.461***	0.391	0.532	0.46***	0.397	0.522
ExplanationBot	0.588***	0.515	0.661	0.589***	0.521	0.658
DelayExplanationBot	0.352***	0.286	0.417	0.363***	0.292	0.435
Constant[#]	0.133	0.081	0.185	0.473	-0.309	1.256
Covariates[+]	No			Yes		
Coefficient comparison	InformBot - ExplanationBot = 0			$\chi^2(1) = 14.84$	Prob > F =	0.000
	InformBot - DelayExplanationBot = 0			$\chi^2(1) = 34.83$	Prob > F =	0.000
	ExplanationBot - DelayExplanationBot = 0			$\chi^2(1) = 7.76$	Prob > F =	0.005

Table 11 DelayExplanationBot and Decision Quality – F1 Score (Study 2)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	0.595***	0.516	0.673	0.611***	0.546	0.676
ExplanationBot	0.793***	0.713	0.874	0.799***	0.727	0.871
DelayExplanatioBot	0.669***	0.596	0.742	0.697***	0.619	0.775
Constant[#]	0.136***	0.078	0.194	-0.025	-0.077	0.028
Covariates[+]	No			Yes		
Coefficient comparison	InformBot - ExplanationBot = 0			$\chi^2(1) = 35.15$ Prob > F = 0.000		
	InformBot - DelayExplanationBot = 0			$\chi^2(1) = 6.59$ Prob > F = 0.000		
	ExplanationBot - DelayExplanationBot = 0			$\chi^2(1) = 6.32$ Prob > F = 0.005		

Note: *N=310. *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of these covariates are significant at p<0.05.*

As shown in Table 12, the self-explanation effect of users in the *DelayExplanationBot* condition had deteriorated by the time the moderation tasks were presented, and the subjects in that condition had a much lower level of error identification (27.4%).

Table 12 Detection of Bot's Error Rate (Study 2)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
ExplanationBot	0.228***	0.123	0.333	0.241***	0.129	0.353
ExplanationDelayBot	-0.12**	-0.213	-0.026	-0.095*	-0.199	0.008
Constant[#]	0.393***	0.322	0.465	0.46	-0.782	1.702
Covariates[+]	No			Yes		
Coefficient comparison	ExplanationBot - MindfulDelayBot = 0			$\chi^2(1) = 35.02$ Prob > chi2 = 0.000		

Note: *N=310. *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'InformBot' condition, in Model 2 it subsumes the 'InformBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of these covariates are significant at p<0.05.*

In terms of the effect of the bot and self-explanation on accuracy of the voluntary moderation behaviors, users who were in the *DelayExplanationBot* treatments had better performance than the *NoBot* group, reporting about 47.6% and 49.7% of the inappropriate comments found in the forum, respectively. The differences in voluntary moderation accuracy levels between the *ExplanationBot*, *InformBot*, and *DelayExplanationBot* groups are all statistically significant.

Table 13 Voluntary Moderation Accuracy (Study 2)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	0.387***	0.320	0.455	0.384***	0.328	0.439
ExplanationBot	0.541***	0.472	0.610	0.542***	0.480	0.605
ExplanationDelayBot	0.408***	0.346	0.471	0.428***	0.371	0.486
Constant[#]	0.089***	0.039	0.138	0.564	-0.476	1.604
Covariates[+]	No			Yes		
Coefficient comparison	InformBot - ExplanationBot = 0			$\chi^2(1)$ = 22.12, Prob > chi2 = 0.0000		
	InformBot - MindfulDelayBot = 0			$\chi^2(1)$ = 10.70, Prob > chi2 = 0.0000		
	ExplanationBot - MindfulDelayBot = 0			$\chi^2(1)$ = 1.88, Prob > chi2 = 0.0000		

Note: *N=310. *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of these covariates are significant at p<0.05.*

In the prior study, I did not find evidence supporting my expectations for overall forum interactions and comment sentiment change. As shown in Table 14, only the subjects in the *DelayExplanationBot* group, who encountered the voluntary moderation task immediately following the self-explanation manipulation, have a higher level of interactions count than the *NoBot* group; there are no significant differences between the subjects in the *NoBot* and the other three groups with respect to comment sentiment change. Similar to the overall interactions count, only the subjects in the *DelayExplanationBot* group had higher voluntary moderation quantity than other groups. Thus, I believe that the subjects in the *ExplanationBot* group had likely depleted their cognitive resources by the time they completed their mandatory rating task and entered the discussion forum for making voluntary contributions. This result, again, provides evidence that the effect of self-generated explanation did not endure multiple tasks (H2).

Table 14 Effect of Self-explanation on Engagement (Study 2)

Panel A: Overall Interactions						
Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformBot	0.959	-1.239	3.157	1.5	-0.761	3.761
ExplanationBot	-1.659	-3.925	0.606	-1.415	-3.730	0.901
DelayExplanationBot	2.105*	0.059	4.151	2.278*	0.082	4.474
Constant[#]	8.508***	6.888	10.128	3.023	-19.938	25.984
Covariates[+]	No			Yes		
Panel B: Comment Sentiment Change						
InformBot	-0.644	-1.862	0.573	-0.441	-1.681	0.799
ExplanationBot	0.525	-0.667	1.718	0.711	-0.559	1.981
DelayExplanationBot	-0.048	-1.176	1.079	-0.08	-1.283	1.122
Constant[#]	-0.222	-0.849	0.404	-2.168	-14.760	10.424
Covariates[+]	No			Yes		
Panel C: Voluntary Moderation Count						
InformBot	-0.998***	-1.507	-0.489	-1.048***	-1.564	-0.532
ExplanationBot	-1.617***	-2.142	-1.092	-1.679***	-2.207	-1.151
DelayExplanationBot	2.155***	1.681	2.629	2.042***	1.541	2.542
Constant[#]	2.571***	2.196	2.947	-1.671	-6.906	3.565
Covariates[+]	No			Yes		

Note: *N=310; *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of them are statistically significant at p<0.05.*

4.0 Study 3: Does accuracy of the bot matter?

In Studies 1-2, I examined cognition as the main attribute of the human agent in the collaborative hybrid designs. In Studies 3 and 4, I examined imperfection as the main attribute of the AI bot agent. In this chapter, I will examine the level of algorithmic imperfection, i.e., accuracy (Wickens and Dixon 2007). Prior research on accuracy and users' adoption of bots' suggestions show that 70% level of accuracy is a threshold, below which users will discard the bot and purely rely on their own assessments (Wickens and Dixon 2007, Yu et al. 2019). Thus, studies examining the behavioral impacts of algorithm accuracy generally chose the accuracy levels above and under 70%. For example, Yin et al. deployed four levels of accuracy to examine the impact of accuracy on trust on automation tools: 60%, 70%, 90%, or 95% (Yin et al. 2019).

As previously discussed, extant research tested only the direct communication approach to inform the users about bots' level of imperfections. With self-generated explanation prompts, users may have higher empathy about bots' limitations, and will be more willing to respond to a less accurate bot (below 70%). Thus, the purpose the Study 3 is to test whether users receiving self-generated explanations, compared to those who were directly informed about the bots' imperfections, could still detect bots' errors and achieve a higher moderation decision quality.

4.1 Methods

The bot accuracy level used in Studies 1 and 2 was 80%. To manipulate the accuracy level in Study 3, I deployed the bot at two lower levels of accuracy: 60%, and 50%. There are three

reasons why I chose two groups with accuracy level less than 70%. First, theoretically, prior study shows that 70% is a threshold below which users no longer rely on the bot (Wickens and Dixon 2007); whereas practically, the Google Perspective API bot, by default, choose 70% as a threshold for classifying whether a comment is toxic or not. Thus, to test whether self-generated explanation can compensate for lower accuracy level, the bot with accuracy lower than 70% are needed. Among many possible accuracy choices below 70%, we set 50% as the bottom line (lowest accuracy) and 60% as another choice (relatively low accuracy). 50% was chosen since modern learning-based algorithms are expected to achieve an accuracy that is more than chance level. 60% is chosen for the convenience of experiment design. we had five comments for subjects to review in the mandatory moderation task, where I needed to assign which comments received an incorrect assessment from the bot. When the bot made one error out of five, the bot accuracy is 80%. When the bot made two errors out of five, the bot accuracy was 60%. There was no way for us to have a design between 60% and 80% accuracy. For the bot with 50% accuracy level, in addition to two wrongly assessed comments, we randomly assigned the last comment to either be marked as appropriate or not appropriate, to get an average accuracy of 50%. For the voluntary moderation task, the bot made correct assessments of 6,8, and 10, to get 50%, 66.7%, and 83.3% levels of accuracy respectively.

To improve individuals' comprehension of bots' function for bots with 50% and 60% accuracy level, I incorporated a new method called "*enhancedExplanation*". In the *explanationBot* group, the same group used in Study 1, users provided one explanation by answering the question "why the bot assessed the comments as appropriate or inappropriate." However, in the e*nhancedExplanation* group, users were prompted to answer an additional question: "Did you

notice any errors the bot made?" This was in addition to answering the initial question about why the bot assessed the comments in a certain way.

4.2 Result

4.2.1 Descriptive Analysis and Manipulation Check

221 subjects (122 males, 98 females, 1 self-identified as 'other') participated in seven experimental conditions focused on bot's accuracy level and enhanced explanation prompt. Subjects' ages varied from 19 to 32 years, with a mean of 20.1. A majority of the subjects (78.73%) were white. 83.71% of the subjects considered themselves as having a little but limited knowledge on the topic of legalized recreational marijuana. Among all the subjects, 184 supported legalization (83.26%), and 37 subjects held the opposite opinion (16.74%). Regarding their online news reading habits, 82.35% regularly read news online, and 79.19% frequently read comments under news articles, sometimes posting a comment on forums. 80.54% of the subjects never or seldom post any comments or replies, and only 51.58% reported having clicked 'thumbs up' or 'thumbs down' buttons during real-world discussion forum interactions. Similar to the participants in Study 1 and Study 2, a majority of the subjects (70.59%) said they never flagged or reported any inappropriate comments during their real-world forum interactions.

In the survey after completing all the tasks in the simulated discussion forum, subjects answered a 5-scale Likert question about how accurate do they think the bot is (1-5 representing very low to very high). The results showed that subjects in the 60% accuracy group perceived higher accuracy than the subjects in the 50% group ($p < 0.05$).

4.2.2 Varying the Bot's Accuracy and Strengthening the Explanation Prompt

In Study 3, I manipulated the bots' accuracy in the presence of regular and enhanced self-explanation prompts. I first compared the cell means of each group to provide model-free evidence for the effect of bots' accuracy and explanation prompt on subjects' decision-making quality. As shown in **Error! Reference source not found.**, Subjects in the *NoBo*t condition, on average, exhibited 62.3% accuracy. Similar to what I found in Study 1, in general, introducing a bot significantly improved subjects' decision quality only if users actively discovered the bot's imperfection. Otherwise, if the bots' accuracy is not high (50% and 60% accuracy level), users' decision-making quality was hurt. Specifically, subjects in the *InformStrictBot* group (the 60% accuracy bot) had 60.6% accuracy, slightly lower than the *NoBot* group, whereas subjects in *InformLenientBot* group (the 50% accuracy bot) had 44.1% accuracy, much lower than the *NoBot* group. Subjects in the explanation groups had higher decision quality than the Inform groups. Among these subjects who were prompted to explain, those working with the strict bot have higher decision quality than the corresponding groups working with the lenient bot.

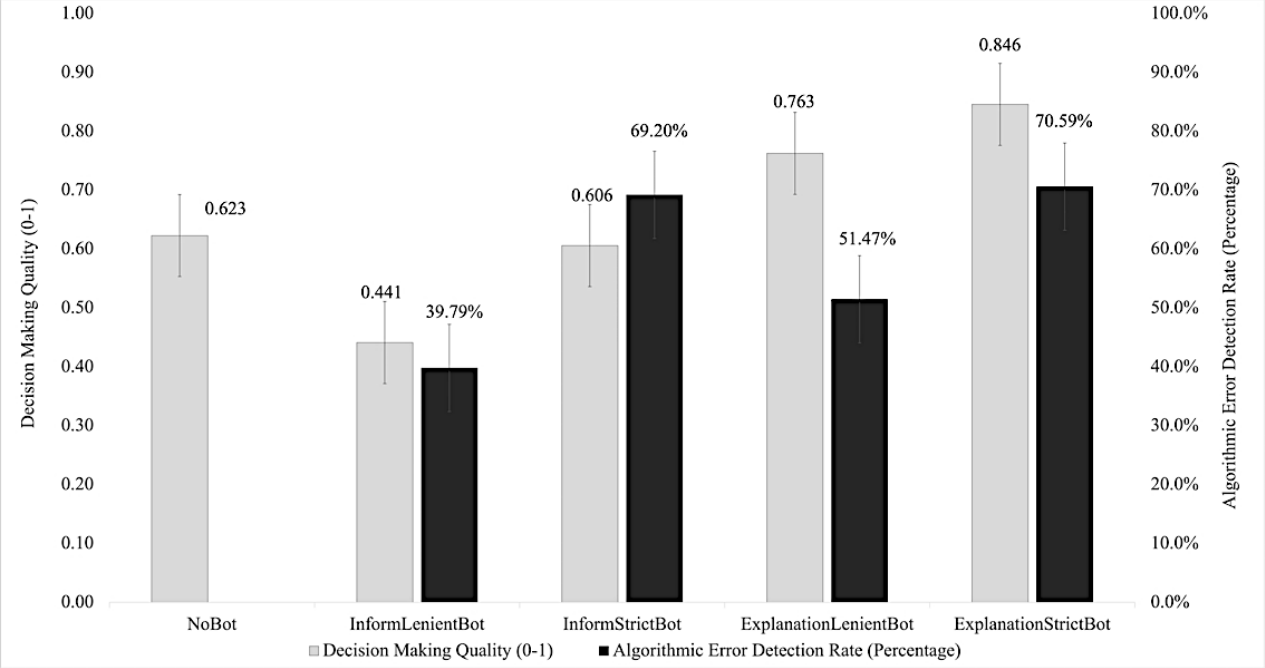


Figure 7 Algorithmic Error Detection Rate and Decision Quality for Mandatory Task (Study 3)

Note. $N = 211$. Error bars indicate 95% confidence interval of the mean

Table 15 provides the corresponding regression results. As shown in Model 1 of Table 15, *NoBot* subjects had a 62.3% accuracy rate. With the bot's assistance, but without active discovery, the accuracy either significantly dropped (*InformLenienetBot*) or was no better than not having the bots' assistance (*InformStrictBot*). However, for both strict and lenient bots, the accuracy of explanation groups increased significantly, relative to the baseline *NoBot* group. The coefficients of explanation groups are significant in the models and the results of the tests comparing the coefficients also reveal statistically significant differences. As presented in Model 2 of Table 15, the results are consistent when the models include a host of pretreatment variables as covariates.

Table 15 Bot and Decision Quality (Study 3)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformStrictBot	-0.017	-0.107	0.073	-0.026	-0.126	0.074
ExplanationStrictBot	0.223***	0.134	0.312	0.224***	0.129	0.319
EnhancedExplanation StrictBot	0.143**	0.049	0.236	0.129*	0.020	0.239
InformLenientBot	-0.182***	-0.271	-0.092	-0.181***	-0.279	-0.083
ExplanationLenientBot	0.140**	0.049	0.231	0.123**	0.030	0.215
EnhancedExplanation LenientBot	0.141*	0.040	0.242	0.127*	0.013	0.242
Constant[#]	0.623***	0.560	0.686	0.000	-0.819	0.819
Covariates[+]	No			Yes		
Coefficient comparison	InformStrictBot - ExplanationStrictBot = 0			$\chi^2(1)$ = 27.86, Prob > chi2 = 0.0000		
	InformStrictBot - EnhancedExplanationStrictBot = 0			$\chi^2(1)$ = 11.20, Prob > chi2 = 0.0010		
	InformStrictBot - InformLenientBot = 0			$\chi^2(1)$ = 12.95, Prob > chi2 = 0.0004		
	InformStrictBot - ExplanationLenientBot = 0			$\chi^2(1)$ = 11.35, Prob > chi2 = 0.0009		
	InformStrictBot – EnhancedExplanationLenientBot = 0			$\chi^2(1)$ = 9.33, Prob > chi2 = 0.0025		
	ExplanationStrictBot – EnhancedExplanationStrictBot = 0			$\chi^2(1)$ = 2.86, Prob > chi2 = 0.0920		
	ExplanationStrictBot - InformLenientBot = 0			$\chi^2(1)$ = 79.25, Prob > chi2 = 0.0000		
	ExplanationStrictBot - ExplanationLenientBot = 0			$\chi^2(1)$ = 3.25, Prob > chi2 = 0.0728		
	ExplanationStrictBot – EnhancedExplanationLenientBot = 0			$\chi^2(1)$ = 2.56, Prob > chi2 = 0.1114		
	EnhancedExplanationStrictBot – InformoutineLenientBot = 0			$\chi^2(1)$ = 46.23, Prob > chi2 = 0.0000		
	EnhancedExplanationStrictBot -ExplanationLenientBot = 0			$\chi^2(1)$ = 0.00, Prob > chi2 = 0.9503		
	EnhancedExplanationStrictBot - EnhancedExplanationLenientBot = 0			$\chi^2(1)$ = 0.00, Prob > chi2 = 0.9719		
	InformLenientBot - ExplanationLenientBot = 0			$\chi^2(1)$ = 47.79, Prob > chi2 = 0.0000		
	InformLenientBot – EnhancedExplanationLenientBot = 0			$\chi^2(1)$ = 39.00, Prob > chi2 = 0.0000		
	ExplanationLenientBot - EnhancedExplanationLenientBot = 0			$\chi^2(1)$ = 0.00, Prob > chi2 = 0.9827		

Note: *N=211. *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of these covariates are significant at p<0.05.*

4.2.3 Explanation Prompt and Detecting Bot Errors

I proceeded to explore the underlying mechanism behind the improvement in the decision quality of subjects in the explanation groups. Results were consistent with the main results in Study 1 when the bot operated at the 60% accuracy level and in the presence of regular and enhanced

self-explanation prompts. Interestingly, when the bot operated at the chance level of accuracy (50%), users did not detect more errors than the *InformLenientBot* group. These results indicate that subjects working with the strict bot were able to detect and rectify much more bot errors than those in the corresponding groups working with the lenient bot.

As shown in Table 16, error detection rates for users who encountered the lenient bot that operated at the chance level of accuracy did not improve even with the explanation prompts. In contrast, when the bot operated at the 60% accuracy level, users were able to detect more bot errors.

Table 16 Detection of Bot's Error Rate (Study 3)

Variables		Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
ExplanationLenientBot		0.012	-0.011	0.035	0.007	-0.024	0.038
EnhancedExplanation LenientBot		0.008	-0.017	0.034	0.010	-0.013	0.034
InformStrictBot		0.029**	0.007	0.052	0.026*	0.006	0.047
ExplanationStrictBot		0.031*	0.008	0.053	0.028*	0.006	0.049
EnhancedExplanationStrictBot		0.021	-0.003	0.045	0.023	0.000	0.047
Constant[#]		0.040***	0.024	0.056	0.386	0.166	0.607
Covariates[+]		No			Yes		
Coefficient comparison	InformStrictBot - ExplanationStrictBot = 0				$\chi^2(1)$ =	0.01, Prob > chi2 =	0.9040
	InformStrictBot - EnhancedExplanationStrictBot = 0				$\chi^2(1)$ =	0.48, Prob > chi2 =	0.4882
	InformStrictBot - ExplanationLenientBot = 0				$\chi^2(1)$ =	2.29, Prob > chi2 =	0.1319
	InformStrictBot – EnhancedExplanationLenientBot = 0				$\chi^2(1)$ =	2.62, Prob > chi2 =	0.1072
	ExplanationStrictBot – EnhancedExplanationStrictBot = 0				$\chi^2(1)$ =	0.66, Prob > chi2 =	0.4161
	ExplanationStrictBot - ExplanationLenientBot = 0				$\chi^2(1)$ =	2.70, Prob > chi2 =	0.1021
	ExplanationStrictBot- EnhancedExplanationLenientBot = 0				$\chi^2(1)$ =	3.01, Prob > chi2 =	0.0844
	EnhancedExplanationStrictBot – ExplanationLenientBot = 0				$\chi^2(1)$ =	0.59, Prob > chi2 =	0.4428
	EnhancedExplanationStrictBot- EnhancedExplanationLenientBot = 0				$\chi^2(1)$ =	0.89, Prob > chi2 =	0.3454
	ExplanationLenientBot- EnhancedExplanationLenientBot = 0				$\chi^2(1)$ =	0.06, Prob > chi2 =	0.8000

Note: *N=211. *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'InformLenientBot' condition, in Model 2 it subsumes the 'InformLenientBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of these covariates are significant at p<0.05.*

4.2.4 The Effects of Bot and Explanation Prompt on User Engagement

In terms of engagement, Study 3's results were consistent with those in Study 1. As the results reported in panels A and B of Table 17 show, I did not find evidence supporting my expectations on overall interactions and comment sentiment: there are no significant differences between the subjects in the *NoBot* and the other groups with respect to their overall interactions and comment sentiment change. The results in the Panel C of Table 17 reveal that the *InformLenientBot* group increased users' intentions to help make voluntary moderation decisions while voluntarily engaging with the discussion forum. This is different than what Study 1 revealed. One possible explanation could be that users who worked with the bot with lower accuracy may have been motivated to help the forum community because they noticed the bot was not working effectively.

Overall, Study 3's results were consistent with the main results in Study 1 when the bot operated at the 60% accuracy level and in the presence of regular and enhanced self-explanation prompts. When the bot operated at the chance level of accuracy (50%), users lost interest in collaborating with the bot, and there were no significant improvements in decision-making quality or voluntary engagement with the forum, even in the presence of enhanced self-explanation prompts. H3 is supported. These results suggest that digital platforms may compensate for the lower accuracy of bots by prompting self-explanation of users, but there is a threshold level of bots' accuracy below which bots will not improve the performance of users. In the experimental context, I determined the threshold level of bot accuracy to be between 50% and 60%, if decision-making quality and user engagement are the main performance outcomes of concern.

Table 17 Effect of Self-explanation, bots' accuracy on Engagement (Study 3)

Panel A: Overall Interactions						
Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformStrictBot	0.523	-3.429	4.474	1.494	-2.145	5.133
ExplanationStrictBot	-0.771	-4.694	3.151	-0.334	-3.729	3.062
EnhancedExplanationStrictBot	2.435	-1.685	6.556	2.294	-1.787	6.374
InformLenientBot	3.787	-0.164	7.739	3.549	-0.493	7.590
ExplanationLenientBot	0.041	-3.972	4.054	-0.281	-4.077	3.514
EnhancedExplanationLenientBot	1.729	-2.736	6.193	0.893	-3.028	4.814
Constant#	7.771***	4.998	10.545	81.230***	43.395	119.065
Covariates^{+}	No			Yes		
Panel B: Comment Sentiment Change						
Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformStrictBot	0.183	-2.237	2.604	-0.354	-3.006	2.298
ExplanationStrictBot	-1.343	-3.746	1.060	-1.893	-4.537	0.751
EnhancedExplanationStrictot	-1.642	-4.167	0.882	-2.281	-5.036	0.474
InformLenientBot	-1.582	-4.002	0.839	-1.797	-4.758	1.164
ExplanationLenientBot	-0.760	-3.218	1.699	-1.036	-3.867	1.795
EnhancedExplanationLenientBot	-0.910	-3.645	1.825	-1.100	-3.605	1.405
Constant#	1.229	-0.471	2.928	4.467	-31.801	40.736
Covariates^{+}	No			Yes		
Panel C: Voluntary Moderation Count						
Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
InformStrictBot	0.608	-0.568	1.784	0.680	-0.452	1.812
ExplanationStrictBot	0.429	-0.739	1.596	0.404	-0.762	1.569
EnhancedExplanationStrictBot	0.797	-0.429	2.023	0.504	-0.829	1.838
InformLenientBot	1.726**	0.550	2.902	1.610**	0.412	2.808
ExplanationLenientBot	0.502	-0.693	1.696	0.547	-0.723	1.816
EnhancedExplanationLenientBot	0.905	-0.424	2.234	0.649	-0.683	1.982
Constant#	1.686*	0.860	2.511	6.341	-5.355	18.037
Covariates^{+}	No			Yes		

Note: *N=221; *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NoBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of them are statistically significant at p<0.05.*

5.0 Study 4: Does valence of the bot matter?

Studies 1 to 3 examined the effects of self-generated explanation and bots' accuracy on moderation decision making quality in the scenarios where users work with a moderation bot that marked inappropriate comments to users. Contrary to my hypothesis H1(b), I failed to find evidence supporting that self-generated explanation will increase user engagement level. One possible explanation is that engagement is more correlated with the atmosphere created by the forum (positive or negative), which can be changed through bots' valence. For example, in the online community context, a positive valence represents the platforms' intention to orient users to appreciate insightful comments and reward others' contributions, rather than to monitor inappropriate comments and punish those who violate the commenting rules (Coursaris et al. 2018). In short, a content moderation bot can either be set up to censor or praise users' comments, which may influence users' engagement. Thus, the purpose of Study 4 was to explore the impact of the bot's valence (positive or negative) on users' engagement.

5.1 Method

5.1.1 Experiment Procedure

The experiment procedures of study 4 are similar to those for the *ExplanationBot* condition in Studies 1-3, except that I varied the bot's processing orientation and user's interaction with the bots was fully voluntary (i.e., not a structured moderation task). In the "*NegativeBot*" condition

the bot marks up comments that violate the forum rules, whereas in the "*PositiveBot*" condition the bot highlights insightful contributions. Similar to Studies 1-3, after reading the article and posting their comments, subjects were shown a pop-up window, which included information about the commenting rules, bot assessment examples, and a prompt to complete the explanation task. After learning about how bot moderates or recommends comments, users were allowed to revisit their own comments and enter the discussion forum. Subjects were able to click "agree" or "disagree" buttons to provide feedback about the bot's assessments, and report or recommend any comments they believed were missed by the bot. Participants were encouraged to engage with the forum as much as they wished, including but not limited to leaving comments, replying to existing comments, up-voting and down-voting comments, reporting any inappropriate comments, and interacting with the bot. In Studies 1-3 and in the *NegativeBot* condition of Study 4, the report-to-the-bot button is used to report inappropriate comments missed by the bot that were discovered by the users. In the *PositiveBot* condition of Study 4, the report button is used to recommend insightful comments that were discovered by users but were not highlighted by the bot.

5.1.2 Manipulations on valence of the bot's processing orientation

I manipulated valence of the bot by changing its processing orientation to derive two experiment groups: (1) *PositiveBot,* and (2) *NegativeBot*. A content moderation bot can either censor or praise users' comments. When the bot's function is to mark or remove comments violating the discussion rules, its valence is negative ("*NegativeBot*"), and when the bot's function is to highlight and recommend insightful contributions to the community, its valence is positive ("*PositiveBot*"). Participants were either told that a bot in the forum would monitor and mark inappropriate comments, or that a bot in the forum would highlight and recommend insightful

comments. After entering the discussion forum, subjects in the *NegativeBot* condition would see a label saying "Bot assessment: not appropriate" attached to suspicious comments. In contrast, subjects in the *PositiveBot* condition would see a label saying "Bot assessment: recommended by the bot" attached to those potentially insightful comments. In both cases, subjects could click "agree with bot" or "disagree with bot" to confirm or correct the bots' assessment.

5.2 Results

5.2.1 Descriptive analysis

In Study 4, 74 subjects (44 males and 30 females) participated in two experimental conditions focused on the valence of the bot's processing orientation. Ages varied from 18 to 26, with a mean of 20.65 years. A majority of the subjects (78.38%) were white. 93.24% of the subjects thought they had some but not a lot of knowledge about legalization of recreational marijuana, and 75.68% of the subjected supported legalization. Regarding their online news reading habits, 86.49% regularly read news online and 82.43% reported having frequently read the discussion forum comments about news articles, but only 24.60% of the subjects had posted their own comments in real-world forums. Similar to the participants in Study 1, 56.76% of the subjects reported having clicked 'thumbs up' or 'thumbs down' during their real-world usage of discussion forums; 87.84% of all the subjects (N=65) said they had never flagged or reported any inappropriate comments in forums in the past.

I compared all experimental conditions to ensure subjects were randomly assigned to the

treatments[1]. Except *Age*, all other coefficients of the pre-treatment variables in the analysis of variance models are not statistically significant (i.e., p-value> 0.05). The significant coefficient of *Age* (p=0.005) indicates some imbalance in the treatment assignment with respect to subjects' age. Further examination of this revealed an outlier subject in Study 4 who was 29 years old, a level substantially higher than the mean *Age* of the sample (20.1). After removing that subject, there are no significant differences between the two treatment groups across all pre-treatment variables[2]. Overall, the results indicate that the randomization of treatment assignment was successful.

5.2.2 Positive Bot Increases Engagement

As hypothesized in H4, I examined whether a positive bot could increase subjects' overall interactions and positivity of their comments while engaging with the discussion forum. As shown in Table 18, subjects in the positive bot condition had higher levels of overall interactions, voluntary moderations, and positive sentiment changes, as compared with those working with the negative bot. These results show strong support for H4, and I conclude that if the valence of the bot is positive, it contributes to a higher level of user engagement and positivity in contributing to the discussion forum.

[1] Cell means of covariates are presented in Table 3 in the Appendix.

[2] I did robustness checks by redoing all the analysis after removing the outlier observation. I found that the exclusion of the outlier observation did not change the results significantly.

Table 18 Bot Valence and Engagement (Study 4)

Panel A: Overall Interactions						
Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
PositiveBot	5.727***	2.800	8.655	4.85*	1.115	8.585
Constant[#]	5.741	3.429	8.053	1.504	-30.376	33.384
Covariates[+]	No			Yes		
Panel B: Comment Sentiment Change						
PositiveBot	2.357***	1.513	3.201	2.248***	0.9415	3.5545
Constant[#]	0.111	-0.242	0.464	-1.738	-12.8891	9.4126
Covariates[+]	No			Yes		
Panel C: Voluntary Moderation Count						
PositiveBot	1.757**	0.482	3.033	1.545*	0.025	3.066
Constant[#]	3.370***	2.354	4.387	1.414	-11.738	14.565
Covariates[+]	No			Yes		

Note: *N=74; *** p < 0.001; ** p < 0.01; * p < 0.05; #: The constant in Model 1 represents the 'NegativeBot' condition, in Model 2 it subsumes the 'NoBot' condition; +: Model 2 includes the following covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of them are statistically significant at p<0.05.*

5.2.3 Exploring Moderation Behavior Differences Between Positive And Negative Bot

In addition, I explored whether the valence of a bot's processing orientation will change users' moderation behaviors. I found that users working with the *PositiveBot* reported more comments to the bot. In other words, subjects in the positive bot group voluntarily recommended more comments to the bot than those reported by the subjects in the *NegativeBot* condition. Those results imply that users are more engaged with the forum when working with a positive bot than a negative bot.

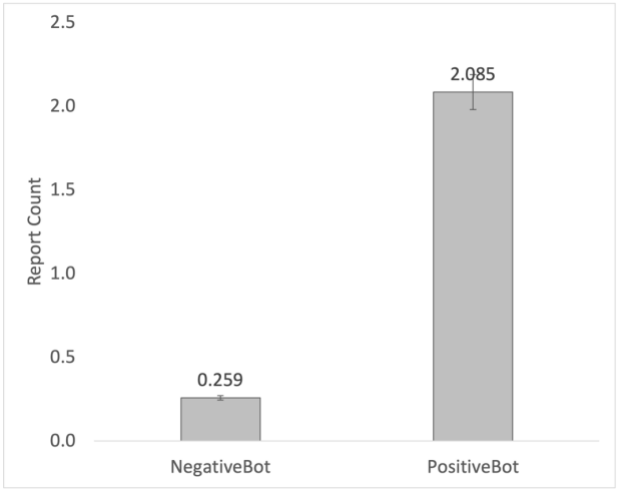


Figure 8 Bot Valence and Reporting (Study 4)

Note. Error bars indicate 95% confidence interval of the mean

5.2.4 Heterogeneous Treatment Effects of Positive Bot on Decision Making Quality

Next, I explored how the effect of a positive bot on how subjects' comment sentiment change varied across the subjects. As shown in Figure 9, the impact of the positive bot varied depending on key features: prior forum interaction habits, voting attitude for the topic, and age. Subjects with limited prior forum interactions (12% of the sample) are more influenced by a positive bot and changed the positivity of their comments more than others. Among those who were more familiar with online discussion forums (88% of the sample), those who supported the legalization of marijuana (66% of the sample) were more likely to be influenced by the bot in improving the positivity of their comments. Among the people who supported legalization of marijuana, subjects younger than 22 (32% of the sample) are more likely to be influenced by the positive bot.

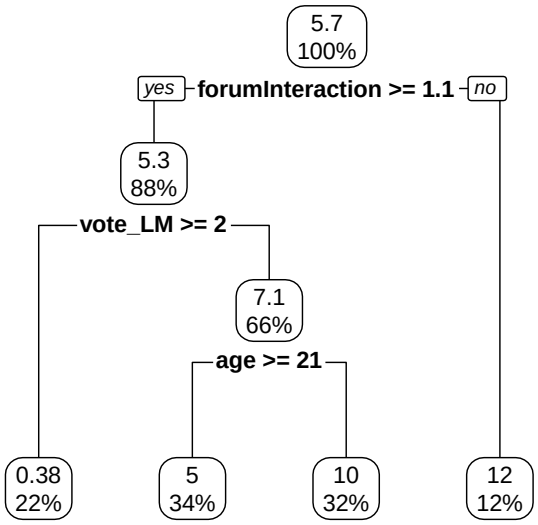


Figure 9 Heterogeneity of the Effect of Positive Bot (Study 4)

Note. The Root Mean Square Error (RMSE) of the causal tree's prediction is 0.58.

6.0 Study 5: Reducing Complacency Potential as Underlying Mechanism

After examining the cognitive attributes of human agents and the imperfect attributes of AI bot agent, I explored the underlying mechanism of the delegation process between these two agents. To compare the differences in complacency potential as the underlying mechanism of direct communication and self-generated explanation groups, I replicated the experiment groups of Study 1 (*NoBo*t, *InformBot*, *ExplanationBo*t) but added a measurement of complacency potential immediately after they expose to bot imperfection illustration (The matrix in Figure 2). In addition, I also added two more groups who also received direct communication about bots' imperfections to observe more variations in the complacency potential. Lastly, to observe the differences in voluntary moderation behaviors and engagement, I decreased the difficulty of the moderated comments to reduce the cognitive depletion level and included more comments in the discussion forum (from 12 to 24 comments) to observe whether the self-generated explanation could also improve the voluntary decision-making quality and engagement on the forum.

6.1 Method

6.1.1 Manipulations

I aim to compare the impact of user-generated explanations of bot actions (*Explanation*) on decision-making quality with three other commonly used approaches that organizations use to communicate bot errors directly to users: (a) a disclaimer stating that the moderation bot can make

mistakes (*Disclaimer*); (b) Materials about the consequences of relying too much on machines (*Alert*); (c) Examples of bot errors with direct pointing (*Inform*). It is important to note that both the *Inform* and *Explanation* groups will encounter specific examples of bot errors. However, while the *Inform* group will be promptly notified of any errors, the *Explanation* group will not be made aware of them until they provide an explanation about the bot actions. All the four groups expose users to bots' potential imperfections either indirectly (i.e., *Explanation* group) or directly (i.e., *Inform*, *Disclaimer,* and *Alert* groups), which may affect their likelihood to detect bots' errors. For example, if users simply rely on the algorithm without questioning, or become aversive to use the algorithm, their algorithmic detection rate will be lower than those who actually pay attention to algorithms' performance.

Based on the four designs to affect users' anticipation about algorithmic errors, I conducted four experimental groups with bot: (1) *DisclaimerBot, (2) AlertBot,* (3) *InformBot,* and (4) *ExplanationBot.* I also included a group without bots' presence to understand the baseline behaviors when no automation is involved: (5) *NoBot* condition. Thus, there are five treatment conditions in Experiment 1. All participants were provided with the rules for commenting on the forum. Those in the bot-assisted conditions were notified that the forum uses a bot for assisting moderation of content.

For *AlertBot* group, I provided subjects an article illustrating three short examples of complacency and the corresponding consequences. The three examples vary from severe and professional settings including an aviation accident, a medical accident due to complacency, and day-to-day life settings including individuals neglecting obvious errors in grammar checking software.

In the *Disclaimer* condition, one sentence saying "The bot is not perfect. Please be prepared

to come across some errors" is used. To provide similar amount of information compared to other conditions, I provided an article on the same news topic of the given reading article (i.e., legalization of recreational marijuana) with similar length to the articles I used in the *AlertBot* condition.

Finally, those who were not assisted by the bot (*NoBot condition*), after learning about the comment rules, were notified that all comments on the forum would be moderated, and they were requested to rate five comments as part of a moderation task before entering the full discussion forum.

6.1.2 Measurements

I followed the same procedures as in prior study but added a measurement for complacency after subjects reading the matrix displaying the bots' error (Figure 3), and right before starting the five-comment moderation tasks. I adopted AICP-R scale to measure the complacency, which reflects subjects' attitude to alleviate workload to the bot and level of lack of responsibility for monitoring the bots (Merritt et al. 2019). There are ten items in the AICP-R (see **Error! Reference source not found.**). The AICP-R is a Likert scale, where the response set for each item ranges from "Strongly disagree" (coded as 1) to "Strongly agree" (coded as 5).

Table 19 Measurement Items for Complacency Potential (adopted from Merritt et al. 2019)

Alleviating Workload to Moderation Bot:

- When I need to review a large volume of comments, it makes sense to delegate the moderation task to a bot.
- When visiting an online forum, if I were looking for certain information, I would let the bot handle some moderation tasks for me.
- The bot should be used to ease users' workload.
- If the bot is available to help me with moderating comments, it makes sense for me to shift more attention to my other activities on the discussion forum (e.g., leave comments and replies to other users, click thumbs up or dislike for threads).
- Distractions and interruptions are less of a problem for me when I have a bot to cover the moderation task.

Lack of Awareness to Monitor the Bot Performance:

- Even if the bot can help me with the moderation task, I should pay attention to its performance.*
- Constantly monitoring the bot's performance is a waste of time.
- Even when I have a lot to do, I am likely to watch the bot carefully for errors.*
- It's not very necessary to pay much attention to the bot when it catches the potential abusive comments on the forum.
- Carefully watching the bot takes time away from more important or interesting things.

**Reversed question.*

6.2 Result

6.2.1 Subjects

I recruited 166 undergraduate students (102 males and 104 females) from a major U.S. institution participated in five experimental conditions focused on bot configuration. Subjects' ages varied from 18 to 29 years, 84.19% of the subjects considered themselves as having some,

but not a lot of knowledge on the topic of legalized recreational marijuana. Majority subjects are regular online social media users. See Appendix B for more detailed descriptions about the subjects.

6.2.2 Manipulation checks

I compared the total time spent on the mandatory content moderation tasks across the experimental groups. As shown in Figure 10, subjects in *DisclaimerBot*, *AlertBot*, and *ExplanationBot* groups spent the most time completing the mandatory moderation tasks (159.879, 179.896, and 163.01 seconds respectively). Nevertheless, subjects were directly told where the bot made mistakes but without explanation manipulations (*InformBot*) spent much less time than the other three with-bot groups, with 89.672 seconds on the moderation task. The cognitive effort that the *InformBot* expended is similar to that of *NoBot* subjects (64.595 seconds). Regression results further corroborated that the level of cognitive effort expended by subjects in the *ExplanationBot*, *DisclaimerBot*, and *AlertBot* conditions are significantly higher than the *InformBot* and *NoBot* groups (see Appendix B Table 1).

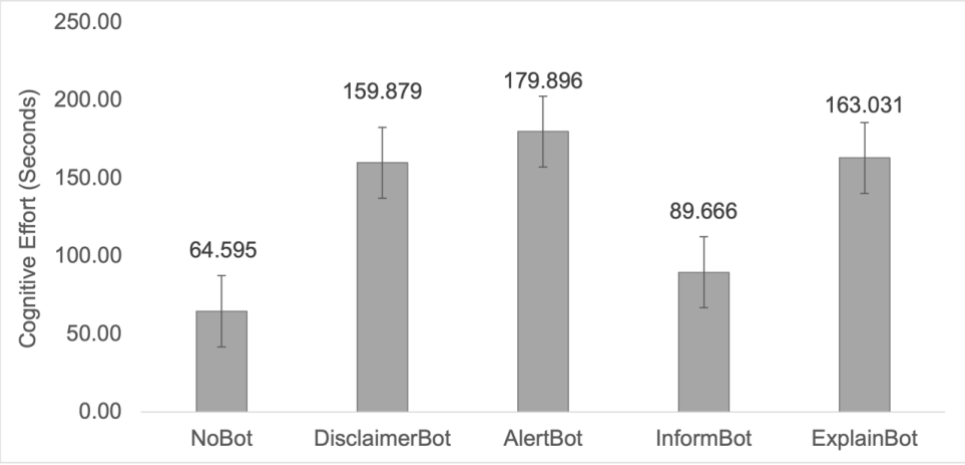


Figure 10 Cognitive Effort on Mandatory Moderation Tasks (Study 5)

Furthermore, I examined the explanation content submitted by subjects in the *ExplanationBot* group. As expected, I found subjects were able to notice the error made by the bot. For example, some subjects were comfortable directly expressing their disagreement with the bot or expressed their confusions about understanding the bots' actions. See Appendix B for some exemplar self-generated explanation contents.

6.2.3 Self-Generated Explanation Improve Error Detection and Decision-Making Quality

Figure 11 shows the model-free evidence for the effect of self-generated explanation on subjects' error detection rate and decision-making quality. I found that the *ExplanationBot* group achieved the highest decision-making quality in both mandatory and voluntary moderation tasks. As shown in Figure 5, subjects who encountered the bot and were prompted to provide explanation (*ExplanationBot*) had a higher level of decision quality (84.4% accuracy) compared to all other groups. The *InformBot* group and *NoBot* group exhibited the lowest level of decision quality (59.3% and 47.1%, respectively). By looking into whether the increased performance of decision quality can be driven by detecting more bots' errors, I found that subjects in *ExplanationBot* were able to detect and rectify 71.85% of the imperfect bot's errors, followed by AlertBot group (57.4%). However, users in the *InformBot* condition only identified 43.7% of the bot's errors, and *DisclaimerBot* group only identified 40.69% bots' errors.

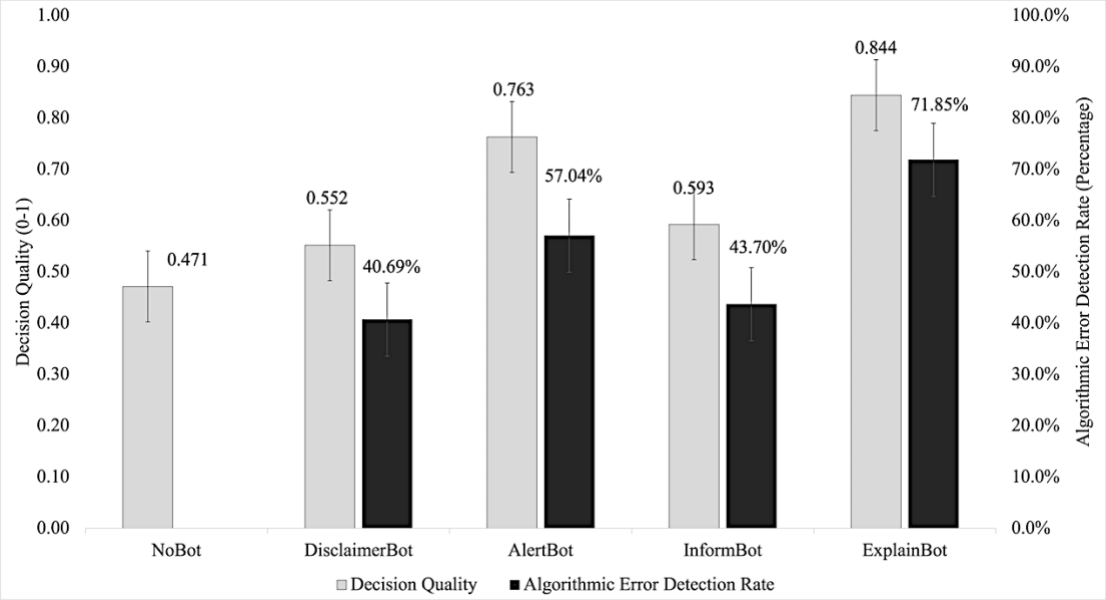


Figure 11 Algorithmic Error Detection Rate and Decision Quality on Mandatory Moderation Task (Study 5)

Table 20 provides the corresponding regression results for testing H1. As shown in Model 1 of Table 20, *NoBot* subjects had a 47.1% accuracy rate. With the bot's assistance, the accuracy of all groups except the *InformBot* groups increased significantly, relative to the baseline *NoBot* group. In particular, the *ExplanationBot* group achieved 37.3% higher accuracy than the *NoBot* group, whose performance is also significantly higher than the *DisclaimerBot* and *AlertBot* groups. The coefficients of *ExplanationBot* are significant in the models and the results of the tests comparing the coefficients also reveal statistically significant differences ($p < 0.001$). As presented in Model 2 of Table 20, the results are consistent when the models include a host of pretreatment variables as covariates. Thus, the results indicate that the explanation practice resulted in higher decision quality, relative to all other groups, which lends strong support to H1. As a robustness check, I also tried other measures of decision quality, and the results are aligned (See Appendix B Table 2).

Table 20 Mandatory Moderation Decision Quality (Study 5)

Variables	Model 1	[95% Conf.	Interval]	Model 2	[95% Conf.	Interval]
DisclaimerBot	0.080	-0.009	0.170	0.061	-0.033	0.155
AlertBot	0.292***	0.200	0.383	0.293***	0.201	0.385
InformBot	0.121**	0.030	0.213	0.109*	0.016	0.201
ExplanationBot	0.373***	0.281	0.465	0.370***	0.276	0.465
Constant	0.471***	0.419	0.524	0.670**	0.204	1.136
Covariates	No			Yes		
Coefficient comparison	Disclaimer - Alert = 0			χ^2=12.48; $p < 0.001$		
	Disclaimer - Inform = 0			χ^2=6.13; $p < 0.05$		
	Disclaimer - Explanation = 0			χ^2=39.75; $p < 0.001$		
	Alert – Inform = 0			χ^2=34.85; $p < 0.001$		
	Alert – Explanation = 0			χ^2=7.42; $p < 0.01$		
	Inform – Explanation = 0			χ^2=74.44; $p < 0.001$		

Note: *N= 166; *** p < 0.001; ** p < 0.01; * p < 0.05; The constant represents the 'NoBot' condition; Covariates: Gender, age, education, race, income, marital status, knowledge level on legalization of marijuana, vote attitude about the legalization of marijuana, the frequency of online news reading, the frequency of online discussion participation (comment, reply, click vote, report any inappropriate contents) and version; none of them are statistically significant at p<0.05.*

I found the same pattern for voluntary moderation decision quality where I calculated the number of correct moderation decisions when users were free to interact with the bot on the discussion forum. For both *forumAccuracySelf* and *forumAccuracy*, I found that the *ExplanationBot* group achieved the highest decision-making quality and bot error detection rates than all other groups. Similar to what I observed in the mandatory moderation task, the *InformBot* group and *NoBot* group, again, exhibited the lowest level of decision quality and bot error detection rates.

6.2.4 Heterogeneous Treatment Effects on Decision Making Quality

As the results in Table 20 demonstrate, the self-generated explanation treatment had the highest positive impact on decision quality. To explore the heterogeneity of this treatment effect on subjects, I analyzed nonparametric causal trees, which are regression trees used for the prediction of treatment effects. Causal trees partition the covariate space into a decision tree that

minimizes prediction error while estimating constant treatment effects within each leaf of the tree (Athey et al. 2017). I used the causal tree method to evaluate how subjects respond differently to the self-generated explanation prompts. Specifically, I used the *causalTree* package in R for analysis and used tenfold cross-validation to avoid overfitting (Athey et al. 2017). As shown in Figure 12, I find that knowledge on the topic and prior forum interaction habits are the most influential factors moderating the impact of explanation treatment on decision making quality. For mandatory tasks (Figure 12, left), I found the impact of explanation depends on topic knowledge and prior forum interaction habits. Specifically, subjects who are not familiar with the topic and had limited forum interaction experience benefitted substantially more from the *ExplanationBot* treatment (their accuracy in classifying messages was 31%).

For voluntary moderation tasks (Figure 12, right), age is another factor in addition to domain knowledge and prior forum interaction habits. Specifically, novice subjects (who had limited domain knowledge and forum interaction experience) younger than 21 benefited the most from the explanation treatment (their accuracy in classifying messages was 33%).

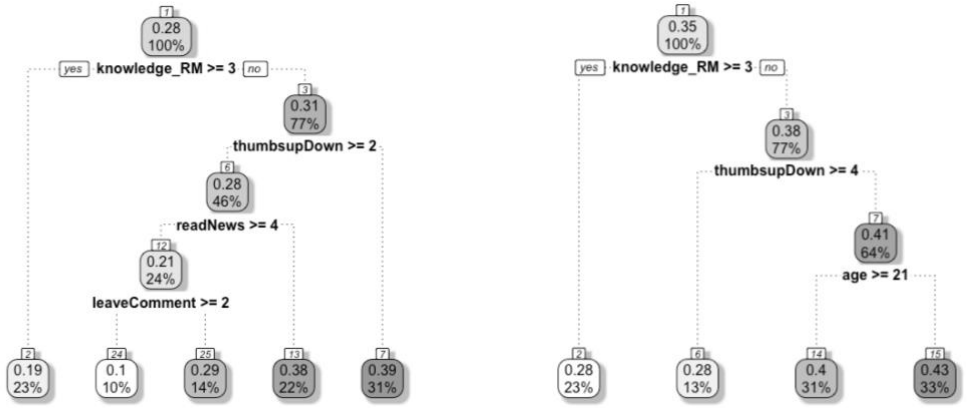
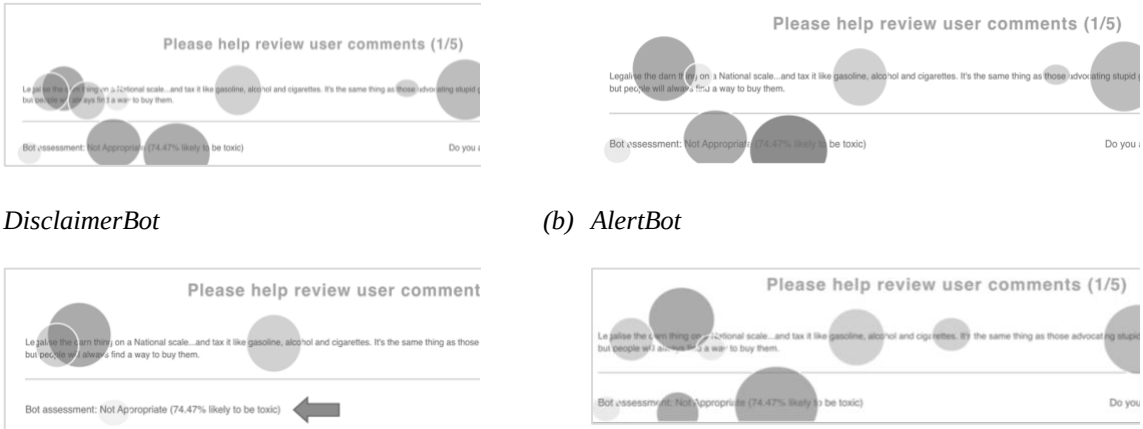


Figure 12 Heterogeneity of the Self-Generated Explanation on Decision Quality (Study 5)

Note: N=166; left: Mandatory task; Right: Voluntary task; Left: causal effect=0.276, Right: causal effect=0.346

To understand where subjects allocated their attention, I compared the heatmaps generated from the mouse movement data across four experiment groups. A heatmap is a graphical representation of data that uses colors to visualize the mouse hovering and staying time (Deng and Poole 2010, Gomez-Marin et al. 2014). Heatmaps can be used to identify trends, patterns, and correlations in the data, and have been widely used to analyze and display social media activity (Ravenscraft 2020).



(a) DisclaimerBot *(b) AlertBot*

(c) InformBot *(d) ExplanationBot*

Figure 13 Heatmaps Generated from Mouse Movement across groups (Study 5)

As shown in Figure 13, the bigger the circle, the more mouse clicks were recorded at the location. The darker the circle, the longer time the mouse stayed. Overall, the items with bigger and darker circles are the items that subjects paid the most and longest attention to. I found subjects in all groups paid attention to the comment to be moderated. There is evidence supporting that subjects in the *Disclaimer*, *Alert*, and *Explanation* treatments paid attention to the bots' advice. However, subjects in *Inform*Bot group barely referred to the bots' advice (see the arrow pointed in the Figure 13 (c)). Please note that I took the heatmap results from one out of five mandatory

moderation tasks for illustration purposes. The results for the rest of the four mandatory content moderation tasks present similar patterns.

6.2.5 Complacency Potential as Potential Underlying Mechanism

I compared the differences in complacency potential across the experiment groups. I took the average of the ten items in the AICP-R, which used a 1-5 Likert scale. The higher the score, the higher complacency level the subject holds for the bot. The *DisclaimerBot* group has the highest level of complacency (3.355), and the *InformBot* group has the lowest level of complacency (2.630). *AlertBot* and *ExplanationBot* groups have similar levels of complacency potential (2.685 and 2.681, respectively). Next, I further examined the two dimensions of complacency potential: (a) the attitude of alleviating workload to automation and (b) the lack of awareness to monitor the automation performance. As shown in Figure 14, I found that the *DisclaimerBot* group displayed a high willingness to delegate workload to the bot (3.862) and lower awareness of monitoring the bots' performance (2.152). With a similarly low level of monitoring the bots' performance awareness (2.230), the willingness of delegating work to the bot of the *InformBot* group is significantly lower than all other groups. As expected, both *ExplanationBot* (3.215) and *AlertBot* group (2.704) perceived a higher level of responsibility to monitor the bots' performance than the other two groups. However, only *ExplanationBot* group remained willing to delegate the moderation work to the bot (3.259) whereas the willingness of alleviating workload to the bot of *AlertBot* group is lower (3.074).

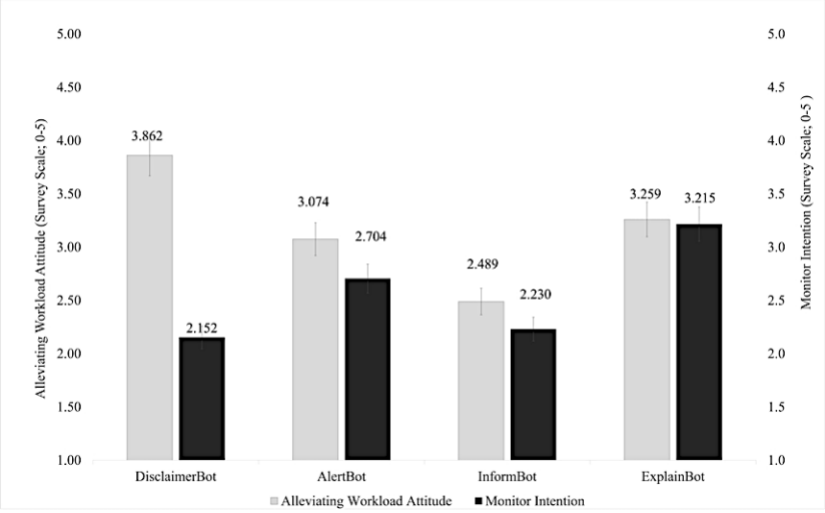


Figure 14 Complacency comparison across group (Study 5)

Note. Error bars indicate 95% confidence interval of the mean

I then assessed whether the main effect of explanation on decision making quality was mediated by reduced complacency potential. I used the causal mediation package for testing and found support for this prediction (Li et al. 2021). As shown in Figure 15, for the mandatory task, self-generated explanation is negatively related with complacency potential, and complacency potential is negatively correlated with decision quality. As shown in the casual mediation analysis results in the Table 21, complacency potential partially mediated the effect of the self-generated explanation on decision quality (14.2% and 33.8% respectively).

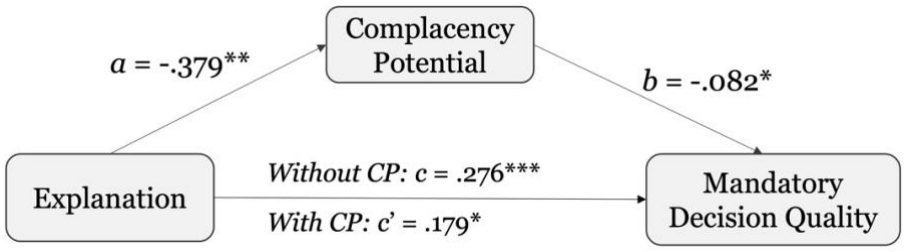


Figure 15 Mediation Effect on Mandatory Decision Quality (Study 5)

Note. $N = 170$. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 21. Causal Mediation analysis results (Mediator: Complacency Potential)

	Mandatory Moderation Quality			Voluntary Moderation Quality		
	Mean	[95% Conf.	Interval]	Mean	[95% Conf.	Interval]
ACME	0.030	0.003	0.067	0.027	0.004	0.059
Direct Effect	0.180	0.084	0.274	0.223	0.142	0.303
Total Effect	0.209	0.118	0.304	0.250	0.173	0.329
% of Total Effect Mediated	0.142	0.097	0.250	0.108	0.082	0.156

Note: N=164. The mediation effect (ACME) is the total effect minus the direct effect.

7.0 Discussion

When collaborating with bots, it is a challenge for users to detect and rectify the bots' errors without incurring algorithm aversion. Errors committed by bots based on machine learning-based algorithms are more unpredictable than those of rule-based algorithms, which exacerbates the challenges of error detection (Jordan and Mitchell 2015). This research demonstrates that a human-AI task design that leverages self-generated explanations of bot actions can help users avoid algorithm aversion and improve decision-making. In our experiments, participants detected more algorithmic errors and achieved higher decision-making quality when they were prompted to explain how an algorithm had arrived at certain conclusions, compared to participants who were directly shown how the algorithm works and its limitations. This generation effect was mediated by a reduction in automation-induced complacency. Finally, even working with an less accurate algorithm, participants who experienced the generation effect achieved higher decision-making quality, although they grew disengaged below a certain threshold of bot accuracy. By comparing four error management training strategies, I found that receiving a simple disclaimer about the bot's imperfections did not prevent users from overreliance on the bot. The users who were informed of the potential risks of overreliance became more alert to the algorithm, and they expelled additional effort to make sense of the bots' suggestions. In contrast, users who were directly informed about the bot's errors became aversive and discarded its recommendations, which hurt their overall decision performance. I summarize my findings in Figure 16.

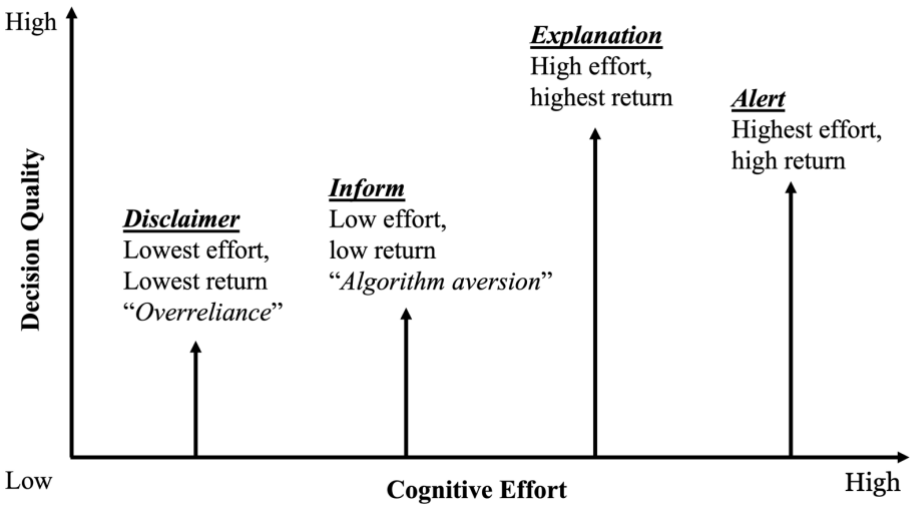


Figure 16 Mediation Effect on Mandatory Decision Quality (Study 5)

7.1 Implications for Researchers Studying Human-AI Hybrid Designs

This paper has explored the human-AI relationship from a collaboration perspective. Quite a few studies on AI have focused on the interaction perspective between humans and bots, or how humans react differently to humans vs. algorithms, but missing the perspective that humans and bots can be team members and complement each other (Chandra et al. 2022, Ebrahimi et al. 2022, Qiu and Benbasat 2014, Riedl et al. 2014, Wang and Benbasat 2016, Yuan and Dennis 2019). We have explored one inhibitory factor for successful human-AI collaboration: end-users' dichotomous expectations about the AI's imperfections, ranging from expecting perfection to having zero tolerance for any errors.

Considering the possibility of humans serving as end-users in the loop, our study contributes to the discussion on improving high quality data input from users for learning-based algorithms (Jordan and Mitchell 2015). Training data quality has an immense impact on the

efficiency, accuracy and complexity of machine learning tasks. Data remains susceptible to errors and irregularities that may be introduced during collection, aggregation or annotation stages. Errors, biases, and data limitations could be addressed by end-users if they were to interact with the AI mindfully and identify any errors (Neff and Nagy 2016). Their corrected input could be used as input to adaptively train the algorithm and improve its accuracy. While researchers and practitioners have focused on improving the quality of models (such as neural architecture search and automated feature selection), efforts towards improving data quality from the end-user perspective have been limited. Our paper is meant as a steppingstone for improving data quality from the end-users, by highlighting the role of self-generated explanations.

To manage users' expectations on bot performance, our findings indicate that researchers should explore ways to trigger users to identify bot imperfections, rather than relying on direct notifications. Most extant training for managing users' expectations on AI treat users as passive information receivers and directly provide them with proceduralized training about bots' imperfections (Montazemi and Wang 2015). Our study findings suggest that passive notifications about bot imperfections do not mitigate overreliance on bots. Instead, researchers should focus on designing interventions that engage users in the error detection process. For example, implementing interactive prompts or challenges that encourage users to critically evaluate bot recommendations and identify potential errors could raise their awareness and vigilance.

Alternative mechanisms for error detection delegation should be investigated, considering the potential for reducing user complacency. Our research highlights the importance of addressing complacency, which can hinder users' ability to detect algorithmic errors. Future research could explore different approaches to delegating error detection responsibilities to users, such as providing feedback mechanisms that require active user input or incorporating gamification

elements to maintain users' engagement and attentiveness. By mitigating complacency, researchers could improve error detection in human-AI hybrid designs.

Lastly, our study highlights that active discovery can be effective even with less accurate bots, increasing the generalizability of the approach. Traditionally, high bot accuracy has been considered a critical factor in user acceptance and reliance. However, this research demonstrates that active user involvement, through self-generated explanations and understanding of the algorithm, can compensate for lower accuracy. Researchers should investigate the conditions under which active discovery can improve error detection, regardless of bot accuracy. This would expand the potential applicability of human-AI hybrid designs to domains where highly accurate algorithms may not be feasible or readily available.

7.2 Implications for Researchers Studying Online Communities

This study addresses moderation behaviors, an understudied type of engagement in online communities (Ray et al. 2014). We recognize moderation behaviors as a valuable form of user engagement in fostering healthy and productive online communities. Researchers studying online communities should consider the role of moderation and explore how algorithms and humans can collaborate to create environments that encourage positive interactions, minimize toxic behavior, and promote a sense of community ownership.

We propose a new approach to managing online discussion involving the collaborative efforts of imperfect algorithms and humans to build and sustain engaging communities. The management of cyberbullying and online harassment in online communities has drawn considerable attention in the IS community (Chan et al. 2019, James et al. 2017, Lowry et al. 2016,

2017, 2019, Matook et al. 2022). Various top-down designs have been developed to control online harassment (Lowry et al. 2017). With a bottom-up approach, the paper demonstrates the possibility of involving users in the process and increasing algorithms' power over the long run, to build a self-organizing community. Instead of relying solely on human moderators or fully automated systems, we suggest a hybrid approach that combines the strengths of both humans and bots. Researchers should investigate methods to integrate imperfect algorithms into the moderation process, allowing them to assist human moderators in identifying and addressing problematic content, while still involving users in the decision-making and error detection processes.

The valence of bots deployed in online communities should not be overlooked, as positive bots tend to increase engagement. Our research findings indicate that deploying positive bots, which provide constructive feedback and support, can promote user engagement in online communities. Researchers studying online communities should explore the effects of bot valence on user behavior, interaction patterns, and community dynamics. By understanding how different bot characteristics influence user engagement, designers can optimize bots' deployment to foster positive and thriving online communities.

7.3 Implications for News Forums and Designers

Crowdsourcing content moderation can be a viable option for news forums and designers when users are provided with bots as assistants and prompted to detect bot errors. News forums and designers can leverage users' collective intelligence by involving them in the moderation process. By integrating bots as assistants and encouraging users to identify and report bot errors, news forums can improve the quality and reliability of user-generated content. Thus, researchers

and designers should investigate strategies for implementing crowdsourced moderation, considering the unique characteristics and challenges of news forums.

A virtuous cycle for self-organizing communities can be created by establishing a feedback loop between vigilant users and efficient bots. This study's findings suggest that an iterative process of user-bot interaction can lead to self-organizing communities. When users detect and report bot errors, the bots can learn and improve, which, in turn, enhances user trust and engagement. News forums and designers should design systems that foster this feedback loop, allowing for the continuous refinement and improvement of both human-AI collaboration and the overall community experience.

Designers should employ diverse methods to involve users while also alerting them of bot imperfections. In doing so, it is crucial for designers to balance users' awareness of bot imperfections with their willingness to collaborate with the bots. Designers should explore different strategies, such as providing informative and transparent explanations of bot limitations, incorporating interactive elements that encourage user input, and designing user interfaces that facilitate users' understanding of how the bots work. By involving users while managing their expectations, designers could create more effective and satisfying human-AI collaborative systems.

Moderation tasks should be distributed based on users' profiles and timing, and considering heterogeneity. To optimize the distribution of moderation tasks, news forums and designers should consider users' profiles, expertise, and availability. By tailoring the allocation of moderation responsibilities to users' characteristics, such as their domain knowledge or their previous performance in error detection, forums could ensure that tasks are assigned to the most qualified individuals. Additionally, considering that users' timing preferences and availability could help

distribute moderation tasks evenly and reduce stress on specific users or time periods, this would promote inclusivity and diversity in the moderation process.

7.4 Limitations and Future Research

7.4.1 The enduring effects of self-generated explanations

Although we found support for higher algorithmic error detection rates when prompting for generation effects via explanation, it would be worthwhile to study whether such effects could endure multiple tasks, and if not, what other interventions could serve the long-term mindful anticipation of errors. Encouraging mindful awareness requires intensive and repeated exercises. One classic and widely cited study (Shapiro et al. 2012) required participants to complete eight weekly, two-hour sessions and attend a half-day retreat (a total of 20 hours). Daily home practice sessions based on audio instructions were accompanied by daily monitoring of both formal and informal meditation practices in a diary. While it would be outlandish to require such intensity in a casual and informal situation like an online community, these practices might be scalable down to a more manageable level with novel platform designs.

7.4.2 The dynamics of human-AI collaboration

One assumption of this paper is that users' error detection behaviors will be consistent in improving the data input quality for iterations of algorithms. However, users may exhibit distinct reactions and behaviors towards a bot in subsequent interactions compared to their initial

interaction. Since our experimental design is cross-sectional rather than longitudinal, we lack evidence to demonstrate the emergence of the hypothesized virtuous cycle. Future research could test the dynamics of human-AI collaboration in a closed cycle, i.e., exploring whether users react differently to an erring bot after their first round of interaction, and assessing whether user input indeed increases algorithms' accuracy, and how such a process could further motivate users to detect algorithmic errors.

7.5 Conclusion

This study has demonstrated the effectiveness of prompting users to provide self-generated explanations about how a bot arrives at certain recommendations as a means to increase their awareness of the bot's imperfections and improve decision-making quality. Human-AI task designs that leverage the generation effect have the potential to reduce automation-induced complacency and improve user engagement with bots for continuous improvement of learning-based systems.

Bibliography

Ackerman R, Thompson VA (2017) Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends Cogn Sci* 21(8):607–617.

Alberdi E, Povyakalo A, Strigini L, Ayton P (2004) Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Acad Radiol* 11(8):909–918.

Alfieri L, Brooks PJ, Aldrich NJ, Tenenbaum HR (2011) Does Discovery-Based Instruction Enhance Learning? *J Educ Psychol* 103(1):1–18.

Apter T (2018) *Passing Judgment: Praise and Blame in Everyday Life.* (WW Norton & Company.).

Atanasov P, Witkowski J, Ungar L, Mellers B, Tetlock P (2020) Small steps to accuracy: Incremental belief updaters are better forecasters. *Organ Behav Hum Decis Process* 160:19–35.

Athey S, Imbens G, Pham T, Wager S (2017) Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*. 278–281.

Bahner JE, Hüper AD, Manzey D (2008) Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *Int J Hum Comput Stud* 66(9):688–699.

Baird A, Maruping LM (2021) The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly* 45(1):315–341.

Baker S (2020) The Boeing 737 Max crashes have revived decades-old fears about what happens when airplane computers become more powerful than pilots. *Business Insider* (February 17) https://www.businessinsider.com/boeing-737-max-fatal-crashes-revive-fears-automation-planes-2020-2.

Bapna S, Benner MJ, Qiu L (2019) Nurturing online communities: An empirical investigation. *MIS Quarterly* 43(2):425–452.

Barlett CP (2017) From theory to practice: Cyberbullying theory and its application to intervention. *Comput Human Behav* 72:269–275.

Barnes CM, Lucianetti L, Bhave DP, Christian MS (2015) "You wouldn't like me when I'm sleepy": Leaders' sleep, daily abusive supervision, and work unit engagement. *Academy of Management Journal* 58(5):1419–1437.

Bateman PJ, Gray PH, Butler BS (2011) The impact of community commitment on participation in online communities. *Information Systems Research* 22(4):841–854.

Baumeister RF (2018) Ego Depletion: Is the Active Self a Limited Resource? *Self-Regulation and Self-Control: Selected Works of Roy F. Baumeister*. 1–387.

Baumeister RF, Gailliot M, DeWall CN, Oaten M (2006) Self-Regulation and Personality: How Interventions Increase Regulatory Success, and How Depletion Moderates the Effects of Traits on Behavior. *J Pers* 74(6):1773–1802.

Becker TE, Klimoski RJ (1989) A field study of the relationship between the organizational feedback environment and performance. *Pers Psychol* 42(2):343–358.

Berton (2018) BGI Cancer. *Tiger Sniff Network* (July 13) https://www.huxiu.com/article/252310.html.

Betsch T, Haberstroh S, Molter B, Glöckner A (2004) Oops, I did it again—relapse errors in routinized decision making. *Organ Behav Hum Decis Process* 93(1):62–74.

Blue V (2017) Google's comment-ranking system will be a hit with the alt-right. *Engadget* https://www.engadget.com/2017-09-01-google-perspective-comment-ranking-system.html.

Boag RJ, Strickland L, Heathcote A, Neal A, Loft S (2019) Cognitive Control and Capacity for Prospective Memory in Complex Dynamic Environments. *J Exp Psychol Gen*.

Bogdanoff A (2015) Saying goodbye to Civil Comments. *Medium*. Retrieved (January 25, 2022), https://medium.com/@aja_15265/saying-goodbye-to-civil-comments-41859d3a2b1d.

Brown SA, Dennis AR, Venkatesh V (2010) Predicting Collaboration Technology Use: Integrating Technology Adoption and Collaboration Research. *Journal of Management Information Systems* 27(2):9–54.

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *J Behav Decis Mak* 33(2):220–239.

Butler BS, Gray PH (2006) Reliability, mindfulness, and information systems. *MIS Q* 30(2):211–224.

Canning EA, Harackiewicz JM (2015) Teach It, Don't Preach It: The Differential Effects of Directly-communicated and Self-generated Utility Value Information. *Motiv Sci* 1(1):47.

Castelo N, Bos MW, Lehmann DR (2019) Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56(5):809–825.

Chan Fung MF chun (2021) Overcoming complacency in the face of infectious disease. *Nature Medicine 2021 27:3* 27(3):363–363.

Chan SH, Song Q, Sarker S, Plumlee RD (2017) Decision support system (DSS) use and decision performance: DSS motivation and its antecedents. *Information and Management* 54(7):934–947.

Chan TKH, Cheung CMK, Wong RYM (2019) Cyberbullying on Social Networking Sites: The Crime Opportunity and Affordance Perspectives. *Journal of Management Information Systems* 36(2):574–609.

Chandra S, Shirish A, Srivastava SC (2022) To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems* 39(4):969–1005.

Chen PY, Hong Y, Liu Y (2018) The Value of Multidimensional Rating Systems: Evidence from a Natural Experiment and Randomized Experiments. *Manage Sci* 64(10):4629–4647.

Cheng J, Bernstein M, Danescu-Niculescu-mizil C, Leskovec J (2017) Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*:1217–1230.

Clemons EK, Gao G, Hitt LM (2006) When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems* 23(2):149–171. https://www.tandfonline.com/action/journalInformation?journalCode=mmis20.

Coe K, Kenski K, Rains SA (2014) Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64(4):658–679.

Coursaris CK, Van Osch W, Albini ACP (2018) Antecedents and Consequents of Information Usefulness in User-generated Online Reviews: A Multi-group Moderation Analysis of Review Valence. *AIS Transactions on Human-Computer Interaction* 10(1):1–25.

Das SR, Chen MY (2007) Yahoo! for amazon: Sentiment extraction from small talk on the Web. *Manage Sci* 53(9):1375–1388.

Davis CJ, Hufnagel EM (2007) Through the eyes of experts: A socio-cognitive perspective on the automation of fingerprint work. *MIS Q* 31(4):681–703.

Degani A (2003) Taming HAL. *Taming HAL*.

Deng L, Poole MS (2010) Affect in web interfaces: A study of the impacts of web page visual complexity and order. *MIS Q* 34(4):711–730.

Dietvorst, Simmons, Massey (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen* 144(1):114–126.

Ebrahimi S, Ghasemaghaei M, Benbasat I (2022) The Impact of Trust and Recommendation Quality on Adopting Interactive and Non-Interactive Recommendation Agents: A Meta-Analysis. *Journal of Management Information Systems* 39(3):733–764.

Endsley MR (2015) Situation awareness: operationally necessary and scientifically grounded. *Cognition, Technology and Work* 17(2):163–167.

Eric J, John W (1986) Effort and accuracy in choice. *Manage Sci* 31(4):395.

Etim B (2017) Why No Comments? It's a Matter of Resources. *New York Times* https://www.nytimes.com/2017/09/27/reader-center/comments-moderation.html.

Evans JSBT (2008) Dual-processing accounts of reasoning, judgment, and social cognition. *Annu Rev Psychol* 59(1):255–278.

Faraj S, Kudaravalli S, Wasko M (2015) Leading Collaboration In Online Communities. *MIS Quarterly* 39(2):393–412.

Farrell S, Lewandowsky S (2000) A Connectionist Model of Complacency and Adaptive Recovery under Automation. *J Exp Psychol Learn Mem Cogn* 26(2):395–410.

FDA (2022) *FDA Warns of Risks Associated with Non-Invasive Prenatal Screening Tests*

Feil-Seifer Maja D, Matari´c M (2011) Automated Detection and Classification of Positive vs. Negative Robot Interactions With Children With Autism Using Distance-Based Features. *HRI'11*.

Feldman S (2019) How Does Facebook Moderate Content. *Statista*. Retrieved (April 28, 2021), https://www.statista.com/chart/17302/facebook-content-moderator/.

Festinger L (1962) Cognitive Dissonance. *Sci Am* 207(4):93–106.

Fischer P, Greitemeyer T, Frey D (2008) Self-Regulation and Selective Exposure: The Impact of Depleted Self-Regulation Resources on Confirmatory Information Processing. *J Pers Soc Psychol* 94(3):382–395.

Fügener A, Grahl J, Gupta A, Ketter W (2021a) Cognitive Challenges in Human-Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research* 33(2).

Fügener A, Grahl J, Gupta A, Ketter W (2021b) WIll humans-in-the-loop become borgs? merits and pitfalls of working with AI. *MIS Q* 45(3):1527–1556.

Galletta DF, Durcikova A, Everard A, Jones BM (2005) Does spell-checking software need a warning label? *Commun ACM* 48(7):82–86.

Galletta DF, Hartzel KS, Susan JE, Jimmie JL, Sandeep R (1996) Spreadsheet Presentation and Error Detection: An Experimental Study. *Journal of Management Information Systems* 13(3):45–63.

Gardiner B, Mansfield M, Anderson I, Holder J, Louter D, Ulmanu M (2016) The dark side of Guardian comments. *The Guardian* https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments?

Gomez P, Ratcliff R, Perea M (2007) A Model of the Go/No-Go Task. *J Exp Psychol Gen* 136(3):389–413.

Gomez-Marin A, Paton JJ, Kampff AR, Costa RM, Mainen ZF (2014) Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nature Neuroscience 2014 17:11* 17(11):1455–1462.

Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45:105681.

Gross D (2014) Online comments are being phased out. *CNN* https://www.cnn.com/2014/11/21/tech/web/online-comment-sections/index.html.

Gu B, Konana P, Rajagopalan B, Chen HWM (2007) Competition among virtual communities and user valuation: The case of investing-related communities. *Information Systems Research* 18(1):68–85.

Gunaratne J, Zalmanson L, Nov O (2018) The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems* 35(4):1092–1120.

Hagger MS, Wood C, Stiff C, Chatzisarantis NLD (2010) Ego Depletion and the Strength Model of Self-Control: A Meta-Analysis. *Psychol Bull* 136(4):495–525.

Halfaker A, Kittur A, Riedl J (2011) Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. *WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration*. 163–172.

Han E, Yin D, Zhang H (2022) Bots with Feelings: Should AI Agents Express Positive Emotion in Customer Service? *Information Systems Research*.

Hardin A, Looney CA, Moody GD (2018) Assessing the Credibility of Decisional Guidance Delivered by Information Systems. *Journal of Management Information Systems* 34(4):1143–1168.

Hattie J, Timperley H (2007) The power of feedback. *Rev Educ Res* 77(1):81–112.

Hawkins RXD, Goodman ND, Goldstone RL (2019) The Emergence of Social Norms and Conventions. *Trends Cogn Sci* 23(2):158–169.

Heart T, Parmet Y, Pliskin N, Zuker A, Pliskin JS (2011) Investigating physicians' compliance with drug prescription notifications. *J Assoc Inf Syst* 12(3):235–254.

Hinojosa AS, Gardner WL, Walker HJ, Cogliser C, Gullifor D (2017) A Review of Cognitive Dissonance Theory in Management Research: Opportunities for Further Development. *J Manage* 43(1):170–199.

Hitron T, Orlev Y, Wald I, Shamir A, Erel H, Zuckerman O (2019) Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. *CHI*. (ACM).

Hollender N, Hofmann C, Deneke M, Schmitz B (2010) Integrating cognitive load theory and concepts of human-computer interaction. *Comput Human Behav* 26(6):1278–1288.

Horne Z, Muradoglu M, Cimpian A (2019) Explanation as a Cognitive Process. *Trends Cogn Sci* 23(3):187–199.

Ilgen DR, Fisher CD, Taylor MS (1979) Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology* 64(4):349–371.

Institute for Safe Medication Practices (2017) *Understanding human over-reliance on technology*

Inzlicht M, Friese M (2019) The Past, Present, and Future of Ego Depletion. *Soc Psychol* 50(5–6):370–378.

James TL, Lowry PB, Wallace L, Warkentin M (2017) The Effect of Belongingness on Obsessive-Compulsive Disorder in the Use of Online Social Networks. *Journal of Management Information Systems* 34(2):560–596.

Jensen ML, Dinger M, Wright RT, Thatcher JB (2017) Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems* 34(2):597–626.

Johns EE, Butler BE (1991) Analysis of Response Time Distributions: An Example Using the Stroop Task. *Psychol Bull* 109(2):340–347.

Johnson SL, Faraj S, Kudaravalli S (2014) Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment. *MIS Quart.* 38(3):795–808.

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science (1979)* 349(6245):255–260.

Joyce E, Kraut RE (2006) Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication* 11(3):723–747.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* 32(3):713–735.

Keenan A (2017) Michigan, Missouri, and Utah legalize marijuana. *Yahoo News* https://www.yahoo.com/news/weed-legal-3-states-heres-expect-new-laws-161924083.html.

Keil FC (2006) Explanation and understanding. *Annu Rev Psychol* 57:227–254.

Keith N, Frese M (2008) Effectiveness of Error Management Training: A Meta-Analysis. *Journal of Applied Psychology* 93(1):59–69.

Kemp S (2023) Digital 2023: Global Overview Report. *Datareportal*. Retrieved (June 30, 2023), https://datareportal.com/reports/digital-2023-global-overview-report.

Khatri V, Samuel BM, Dennis AR (2018) System 1 and System 2 cognition in the decision to adopt and use a new technology. *Information and Management* 55(6):709–724.

Kluger AN, DeNisi A (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull* 119(2):254–284.

Lai V, Tan C (2019) On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*:29–38.

Lambrecht A, Tucker C (2017) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Manage Sci* 65(7):2966–2981.

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.

Li X, Grahl J, Hinz O (2021) How Do Recommender Systems Lead to Consumer Purchases? A Causal Mediation Analysis of a Field Experiment. *Information Systems Research* 33(2):620–637.

Liquin E, Lombrozo T (2017) Explain, Explore, Exploit: Effects of Explanation on Information Search. *CogSci*.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organ Behav Hum Decis Process* 151:90–103.

Lomas N (2015) Can Civil Comments Kill The Internet Troll? *TechCrunch*. Retrieved (January 25, 2022), https://techcrunch.com/2015/10/31/can-civil-comments-kill-internet-trolls/?guccounter=1.

Lombrozo T (2006) The structure and function of explanations. *Trends Cogn Sci* 10(10):464–470.

Lombrozo T, Carey S (2006) Functional explanation and the function of explanation. *Cognition* 99(2):167–204.

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46(4):629–650.

Lowry PB, Moody GD, Chatterjee S (2017) Using IT Design to Prevent Cyberbullying. *Journal of Management Information Systems* 34(3):863–901.

Lowry PB, Zhang J, Moody GD, Chatterjee S, Wang C, Wu T (2019) An Integrative Theory Addressing Cyberharassment in the Light of Technology-Based Opportunism. *Journal of Management Information Systems* 36(4):1142–1178.

Lowry PB, Zhang J, Wang C, Siponen M (2016) Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the

social structure and social learning model. *Information Systems Research* 27(4):962–986.

Mahlfeld W, Hasse C, Grasshoff D, Bruder C (2011) The effect of complacency potential on human operators' monitoring behaviour in aviation. Waard D de, Gérard N, Onnasch L, Wiczorek R, Manzey D, eds. *Human Centred Automation*. (Shaker Publishing, Maastricht, the Netherlands), 1–12.

Maltz M, Shinar D (2004) Imperfect in-vehicle collision avoidance warning systems can aid drivers. *Hum Factors* 46(2):357–366.

Margetts H, Dorobantu C (2019) Rethink government with AI. *Nature* 568(7751):163–165.

Marvin CB, Shohamy D (2016) Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *J Exp Psychol Gen* 145(3):266–272.

Marvin R (2019) How Google's Jigsaw Is Trying to Detoxify the Internet. *PC Mag* https://www.pcmag.com/news/how-googles-jigsaw-is-trying-to-detoxify-the-internet.

Matook S, Dennis AR, Wang YM (2022) User Comments in Social Media Firestorms: A Mixed-Method Study of Purpose, Tone, and Motivation. *Journal of Management Information Systems* 39(3):673–705.

McAllister CP, Mackey JD, Perrewé PL (2018) The role of self-regulation in the relationship between abusive supervision and job tension. *J Organ Behav* 39(4):416–428.

McCurdy MP, Viechtbauer W, Sklenar AM, Frankenstein AN, Leshikar ED (2020) Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychon Bull Rev* 27(6):1139–1165.

McKnight DH, Liu P, Pentland BT (2020) Trust Change in Information Technology Products. *Journal of Management Information Systems* 37(4):1015–1046.

Merritt SM, Ako-Brew A, Bryant WJ, Staley A, McKenna M, Leone A, Shirase L (2019) Automation-induced complacency potential: Development and validation of a new scale. *Front Psychol* 10(FEB):225.

Metcalfe J (2017) Learning from Errors. *Annu Rev Psychol* 68:465–489.

Metzger U, Parasuraman R (2005) Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Hum Factors* 47(1):35–49.

Millman Z, Hartwick J (1987) The impact of automated office systems on middle managers and their work. *MIS Q* 11(4):479–490.

Montazemi AR, Wang S (2015) The Effects of Modes of Information Presentation on Decision-Making: A Review and Meta-Analysis. *Journal of Management Information Systems* 5(3):101–127.

Moon JY, Sproull LS (2008) The role of feedback in managing the internet-based volunteer work force. *Information Systems Research* 19(4):494–515.

Moray N, Inagaki T, Itoh M (2000) Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *J Exp Psychol Appl* 6(1):44–58.

Mosier K, Skitka L (1996) Human Decision Makers and Automated Decision Aids: Made for Each Other? Parasuraman R, Mouloua M, eds. *Automation and Human Performance: Theory and Applications*. (CRC Press), 201–220.

Muraven M (2012) Ego depletion: Theory and evidence. Deci EL, Ryan M, eds. *Oxford library of psychology. The Oxford handbook of human motivation*. (Oxford University Press.), 111–126.

Muraven M, Baumeister RF (2000) Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle? *Psychol Bull* 126(2):247–259.

Muraven M, Buczny J, Law KF (2019) Ego depletion: Theory and evidence. Ryan. R. M., ed. *The Oxford handbook of human motivation*. (Oxford University Press), 113–134.

Naquin CE, Kurtzberg TR (2004) Human reactions to technological failure: How accidents rooted in technology vs. human error influence judgments of organizational accountability. *Organ Behav Hum Decis Process* 91:129–141.

Nass C, Steuer J, Tauber E (1994) Computers are Social Actors. *CHI '94*.

Neff G, Nagy P (2016) Talking to Bots: Symbiotic Agency and the Case of Tay. *Int J Commun* 10(0):17.

Nielsen FÅ (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CEUR Workshop Proc*. 93–98.

Norris-Watts C, Levy PE (2004) The mediating role of affective commitment in the relation of the feedback environment to work outcomes. *J Vocat Behav* 65(3):351–365.

Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437(7063):1291–1298.

Oestreicher-singer G, Zalmanson L (2013) Content or Community? A Digital Business Strategy for Content Providers in the Social Age. *MIS Quart.* 37(2):591–616.

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. *Hum Factors* 52(3):381–410.

Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. *Hum Factors* 39(2):230.

Park EH, Werder K, Cao L, Ramesh B (2022) Why do Family Members Reject AI in Health Care? Competing Effects of Emotions. *Journal of Management Information Systems* 39(3):765–792.

Park JS, Barber R, Kirlik A, Karahalios K (2019) A slow algorithm improves users' assessments of the algorithm's accuracy. *Proc ACM Hum Comput Interact* 3(CSCW):15.

Pavlou PA, Dimoka A (2006) The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums and seller differentiation. *Information Systems Research* 17(4):392–414.

Qiu L, Benbasat I (2014) Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems* 25(4):145–182.

Ragupathi M, Haas EM (2011) Designing a robotic colorectal program. *J Robot Surg* 5(1):51–56.

Rai A, Constantinides P, Sarker S (2019) Next-Generation Digital Platforms: Toward Human–AI Hybrids. *MIS Quarterly* 43(1):iii–ix.

Ransbotham S, Kane GC (2011) Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quart.* 35(3):613–627.

Ravenscraft E (2020) Almost Every Website You Visit Records Exactly How Your Mouse Moves. *Medium* (February 4) https://onezero.medium.com/almost-every-website-you-visit-records-exactly-how-your-mouse-moves-4134cb1cc7a0.

Ray S, Kim SS, Morris JG (2014) The central role of engagement in online communities. *Information Systems Research* 25(3):528–546.

Renner N (2016) As sites abandon comments, The Coral Project aims to turn the tide. *Columbia J Rev*.

Riedl R, Mohr P, Kenning P, Davis F, Heekeren H (2014) Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. *Journal of Management Information Systems* 30(4):83–114.

Roscoe RD, Chi MTH (2008) Tutor learning: The role of explaining and responding to questions. *Instr Sci* 36(4):321–350.

Rosen CC, Levy PE, Hall RJ (2006) Placing perceptions of politics in the context of the feedback environment, employee attitudes, and job performance. *Journal of Applied Psychology* 91(1):211–220.

Sandberg J, Holmström J, Lyytinen K (2020) Digitization and Phase Transitions in Platform Organizing Logics: Evidence from the Process Automation Industry. *MIS Quarterly* 44(1).

Sarker Saonee, Valacich JS, Sarker Suprateek (2005) Technology Adoption by Groups: A Valence Perspective *. *J Assoc Inf Syst* 6(2):37–71.

Schmeichel BJ, Vohs KD, Baumeister RF (2003) Intellectual Performance and Ego Depletion: Role of the Self in Logical Reasoning and Other Information Processing. *J Pers Soc Psychol* 85(1):33–46.

Schuetzler RM, Grimes GM, Scott Giboney J (2020) The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems* 37(3):875–900.

Sebok A, Wickens CD (2017) Implementing Lumberjacks and Black Swans Into Model-Based Tools to Support Human-Automation Interaction. *Hum Factors* 59(2):189–203.

Shapiro SL, Jazaieri H, Goldin PR (2012) Mindfulness-based stress reduction effects on moral reasoning and decision making. *Journal of Positive Psychology* 7(6):504–515.

Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Manage Sci* 59(6):1425–1443.

Simester D, Timoshenko A, Zoumpoulis SI (2020) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Manage Sci* 66(6):2495–2522.

Singh IL, Molloy R, Parasuraman R (1993) Automation-Induced "Complacency": Development of the Complacency-Potential Rating Scale. *Int J Aviat Psychol* 3(2):111–122.

Skitka LJ, Mosier KL, Burdick M, Rosenblatt B (2009) Automation Bias and Errors: Are Crews Better Than Individuals? *Int J Aviat Psychol* 10(1):85–97.

Slamecka NJ, Graf P (1978) The generation effect: Delineation of a phenomenon. *J Exp Psychol Hum Learn* 4(6):592–604.

Srinivasan R, Len Sarial-Abi G (2021) When Algorithms Fail: Consumers' Responses to Brand Harm Crises Caused by Algorithm Errors. *J Mark* 85(5):74–91.

Srivastava R, Bharti PK, Verma P (2022) Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. *IJACSA) International Journal of Advanced Computer Science and Applications* 13(3).

Sternberg RJ (2000) Images of mindfulness. *Journal of Social Issues* 56(1):11–26.

Tarantola A (2017) New York Times picks an AI moderator over a Public Editor. *Engadget* https://www.engadget.com/2017-05-31-new-york-times-picks-an-ai-moderator-over-a-public-editor.html.

Thomas Z (2020) Facebook content moderators paid to work from home. *BBC* https://www.bbc.com/news/technology-51954968.

Trusov M, Bucklin RE, Pauwels K (2009) Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *J Mark* 73(5):90–102.

Tversky A, Kahneman D (1974) Judgment under Uncertainty: Heuristics and Biases. *Science (1979)* 185(4157):1124–1131.

Urban GL, Hauser JR (2004) "Listening In" to Find and Explore New Combinations of Customer Needs. *J Mark* 68(2):72–87.

Valorinta M (2009) Information technology and mindfulness in organizations. *Industrial and Corporate Change* 18(5):963–997.

Wang W, Benbasat I (2014) Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems* 23(4):217–246.

Wang W, Benbasat I (2016) Empirical Assessment of Alternative Designs for Enhancing Different Types of Trusting Beliefs in Online Recommendation Agents. *Journal of Management Information Systems* 33(3):744–775.

Wickens CD, Dixon SR (2007) The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theor Issues Ergon Sci* 8(3):201–212.

Willemsen LM, Neijens PC, Bronner F, de Ridder JA (2011) “Highly Recommended!” The Content Characteristics and Perceived Usefulness of Online Consumer Reviews. *Journal of Computer-Mediated Communication* 17(1):19–38.

Williams JJ, Lombrozo T (2010) The role of explanation in discovery and generalization: Evidence from category learning. *Cogn Sci* 34(5):776–806.

Williams JJ, Lombrozo T (2013) Explanation and prior knowledge interact to guide learning. *Cogn Psychol* 66(1):55–84.

Yin M, Vaughan JW, Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*.

Yoshida N, Yonezawa T (2018) Arousal and Valence in Ro-bot’s Emotional Expression of, Breathing and Heartbeat. *HAI’18*.

You S, Yang CL, Li X (2022) Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *JOURNAL OF MANAGEMENT INFORMATION SYSTEMS* 39(2):336–365.

Yu K, Berkovsky S, Taib R, Zhou J, Chen F (2019) Do I trust my machine teammate? An investigation from perception to decision. *International Conference on Intelligent User Interfaces, Proceedings IUI* Part F147615:460–468.

Yuan L (Ivy), Dennis AR (2019) Acting Like Humans? Anthropomorphism and Consumer’s Willingness to Pay in Electronic Commerce. *Journal of Management Information Systems* 36(2):450–477.

Zhang Tao, Kaber David B, Biwen ·, Manida Z·, Prithima Mosaly S·, Hodge L, Zhang T, et al. (2010) Service robot feature design effects on user perceptions and emotional responses. *Intelligent Service Robotics 2010 3:2* 3(2):73–88.

Zhao B, Olivera F (2006) Error reporting in organizations. *Academy of Management Review* 31(4):1012–1030.

Zhu H, Zhang A, He J, Kraut RE, Kittur A (2013) Effects of peer feedback on contribution: A field experiment in Wikipedia. *Conference on Human Factors in Computing Systems - Proceedings*. 2253–2262.