Can AI Enhance People’s Support for Online Moderation and Their Openness to Dissimilar Political Views?

Magdalena Wojcieszak, Arti Thakur, João Fernando Ferreira Gonçalves, Andreu Casas, Ericka Menchen-Trevino, & Miriam Boon

1Department of Communication, University of California, Davis, CA 95618, USA
2Amsterdam School of Communication Research, University of Amsterdam, 1018 WV Amsterdam, The Netherlands
3Department of Communication, University of California, Davis, CA 95618, USA
4Department of Media & Communication, Erasmus University Rotterdam, Rotterdam, 3000 DR, The Netherlands
5Department of Communication Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
6School of Communication, American University, Washington, DC 20016, USA
7Amsterdam School of Communication Research, University of Amsterdam, 1018 WV Amsterdam, The Netherlands

Although artificial intelligence is blamed for many societal challenges, it also has underexplored potential in political contexts online. We rely on six preregistered experiments in three countries (N = 6,728) to test the expectation that AI and AI-assisted humans would be perceived more favorably than humans (a) across various content moderation, generation, and recommendation scenarios and (b) when exposing individuals to counter-attitudinal political information. Contrary to the preregistered hypotheses, participants see human agents as more just than AI across the scenarios tested, with the exception of news recommendations. At the same time, participants are not more open to counter-attitudinal information attributed to AI rather than a human or an AI-assisted human. These findings, which—with minor variations—emerged across countries, scenarios, and issues, suggest that human intervention is preferred online and that people reject dissimilar information regardless of its source. We discuss the theoretical and practical implications of these findings.

Lay Summary

In the era of unprecedented political divides and misinformation, artificial intelligence (AI) and algorithms are often seen as the culprits. In contrast to these dominant narratives, we argued that AI might be seen as being less biased than a human in online political contexts. We relied on six preregistered experiments in three countries (the United States, Spain, Poland) to test whether internet users perceive AI and AI-assisted humans more favorably than simply humans; (a) across various distinct scenarios online, and (b) when exposing people to opposing political
information on a range of contentious issues. Contrary to our expectations, human agents were consistently perceived more favorably than AI except when recommending news. These findings suggest that people prefer human intervention in most online political contexts.

**Keywords:** Artificial Intelligence, AI, algorithms, content moderation, news recommendations, polarization, biased information processing, social media, counter-attitudinal views, news, bias, online moderation, perceived justice

doi:10.1093/jcmc/zmab006

**Can AI enhance people’s support for online moderation and their openness to dissimilar political views?**

Political divides, populism, and misinformation are on the rise. Many observers attribute these problems to the nature of online discussions and—more specifically—to algorithms and artificial intelligence (AI). These technologies are blamed for minimizing diverse exposure and placing users in echo chambers (Pariser 2011; Sunstein 2009) and automated bots for populating online discussions with incivility (Coleman, 2018) and spreading low-quality content (Woolley & Howard, 2016). Although these worries are contradicted by other work (e.g., user preferences matter more than algorithms to congenial exposure; Baksy, Messing, & Adamic, 2015; humans spread misinformation at greater rates than bots; Vosoughi, Roy, & Aral, 2018; algorithmic news recommendations do not limit diverse exposure; Möller et al., 2018), algorithms and AI are blamed for pressing political problems.

We argue that AI has important democratic potential that is understudied in political contexts online. Increasingly, AI is used to address some of the ills of online discussions and hoped to solve some problems (Zuckerberg, 10 April 2018). For instance, social media platforms rely on AI and/or on its partnership with human moderators to monitor hate speech, and media organizations use AI for news recommendations or comment moderation (Ofcom, 2019; Roberts, 2020; Thurman et al., 2019). And yet, despite this growing relevance of AI and AI-human partnership, limited work compares users’ reactions to AI and AI-assisted humans, versus humans (Tandor et al., 2020; Waddell, 2019a; Wölker & Powell, 2018), and—to our knowledge—no studies systematically compare users’ reactions to these three agents across multiple political contexts online.

This project fills this gap by addressing two fundamental questions: (a) Do people see discussion moderation, content generation, and news recommendation decisions as more just when these are made by AI, an AI-assisted human, or a person? And (b) are people more open to cross-cutting political information when it is attributed to AI rather than a human? Our overarching theoretical expectation is that—in political contexts online—AI and AI-assisted humans will have an advantage over human agents. After all, AI has no apparent political affiliation and no clear self-interest, and thus should be perceived as fair and unbiased (Sundar, 2008), as detailed below.

We rely on six preregistered experiments embedded in three studies in distinct countries (the United States, Spain, Poland; total \(N = 6,728\)). In Experiment 1, we examine whether attributing content moderation, recommendation, and generation decisions to AI, an AI-assisted human, or a human influences perceived justice (i.e., seeing the agent as fair and politically objective, and the decision as legitimate and fair; Lee, 2018; Ötting & Maier, 2018). We explore these effects across four situations, in which AI is or could be used. In Experiment 2, we test whether attributing political content to AI, an AI-assisted human, or a human makes people more open to counter-attitudinal
information (i.e., agreement and perceived quality; the source’s credibility and objectivity) across three distinct issues.

Contrary to the preregistered hypotheses, participants evaluated the agent and its decision as more just when it was attributed to a human than to an AI (Experiment 1). In contrast, with a few exceptions, participants rejected counter-attitudinal information uniformly and regardless of its source (Experiment 2). These findings, which emerged across countries, scenarios, and issues, suggest that users prefer human content moderation, generation, and recommendation online, and that AI has some, but limited, potential to attenuate partisan biases triggered by cross-cutting exposure.

Before outlining our theoretical foundations, we note that substantial terminological ambiguity exists in past work (Wang, 2019). An algorithm is typically understood as “encoded procedures for transforming input data into a desired output, based on specified calculations” (Gillespie 2014, p. 1). Thus, the term can encompass computational applications in a variety of fields and contexts. AI, in turn, refers to the capability of a machine to exhibit human-like performance at a task and to computations developed by machines that are contrasted with human performance (Ofcom, 2019). As our project compares machine and human performance, we use the term AI to refer to the nonhuman agents performing what are typically human tasks. Other terms, such as automated decision-makers, could be used. However, due to the “hype” surrounding AI, the wide use of the term, and AI’s portrayal in popular fiction (Sundar, Waddell, & Jung, 2016), “Artificial Intelligence” is likely to be more easily understood when referring to algorithmic performance as a counterpart to human performance. As such, this term was used in our project.

The role of AI in political contexts

Growing numbers of citizens get their news and discuss politics online. Roughly 80% of Europeans turn to the internet for information (Newman, Fletcher, Levy, & Nielsen, 2015), 79% of Americans read or contribute comments to online news (Stroud, Scacco, & Curry, 2016), and millions discuss politics on social media (Barberá, et al., 2019). Scholars and observers worry that using new media for news and discussion presents problems, such as misinformation spread (Vosoughi et al., 2018), incivility, and hate speech (Coe, Kenski, & Rains, 2014; Kim & Kim, 2019; Stroud et al., 2016), or political discussions among like-minded users (Barberá, et al., 2015). For instance, an EU report states that “Online platforms facilitate the high-speed, large-scale and targeted spreading of conspiracy theories, disinformation and junk news . . . (leading to) polarised, emotional debates in which users are automatically fed information confirming existing cognitive biases” (Bentzen, 2018, p. 3).

Many social media platforms and news organizations rely on human moderators to address these problems by supervising users or controlling conversations to align them with community standards. Reddit moderators screen posts to delete malicious content and penalize users for abusive behavior (Almerekhi, Kwak, & Jansen, 2020) and journalists or community managers tame uncivil discussions and interact with problematic commenters (Masullo, Riedl, & Huang, 2020). Such human intervention, however, is not scalable, takes a psychological toll on the moderators, subjects them to harassment, and is questioned by users, who see moderators’ decisions as biased or illegitimate (Almerekhi et al., 2020; Riedl, Masullo, & Whipple, 2020; Shen, Miller, Jo, & Rose, 2018).

Therefore, AI increasingly assists human moderators. It filters which unclear or nuanced content moderators need to evaluate and offers information about the cases to be reviewed. This cooperation aids human moderators, who would otherwise have to view all the content reported by users (e.g., violent extremism), helps to further train AI to improve its accuracy, and offers oversight of the decisions
of AI. After all, AI does not capture intent and its contextual understanding is limited, which may lead to errors. Consider the cases of Facebook removing an image of a statue of Neptune in Italy for being sexually explicit or the iconic photo of a girl fleeing a napalm bombing in Vietnam for featuring child nudity. AI also assists journalists in data retrieval and integration, sourcing articles, and generating stories, allowing newsrooms to operate more cost-effectively.

As AI continues to advance, growing more able to understand the context and interactive discussions, it may shift from an auxiliary to a primary role. As some claim, “rapid progress in the field of artificial intelligence has endowed algorithms with the abilities to understand and produce natural language, learn from experience, and even understand and mimic human emotions” (Castelo, Bos, & Lehmann, 2019, p. 810, b). AI not only nudges users toward certain behaviors but also makes autonomous decisions. It is used by social media platforms to automatically take down content rated as false by fact-checkers, among other actions (Facebook, 2020), and by news media to recommend news to users (Möller, et al., 2018) or to publish news independently (e.g., The Guardian, 8 September 2020).

**Theorizing and evidence on users’ reactions to AI**

Given the growing relevance of AI across political contexts online, research on how users perceive AI tools and their decisions is needed. Existing theories offer conflicting predictions and findings are inconclusive. The Computers are Social Actors (CASA, Nass, Steuer, & Tauber, 1994) framework suggests that people react to machines as they do to humans, apply the same standards and expectations, and treat machines as if they had their own intentions (Reeves & Nass, 1998). It follows that there should be no difference in how users perceive AI versus a human. Yet, the rapid sociotechnological changes (i.e., human–technology interactions are more prevalent and elaborate) reveal limitations in the original CASA model and some recent research does not support it (Gambino, Fox, & Ratan, 2020 for a review).

In contrast, the algorithm aversion literature argues that users distrust algorithmic output for various reasons (e.g., algorithms are seen as not transparent, do not possess human traits). Thus, people should not only have low trust in AI but also prefer a human over AI, especially for decisions traditionally made by humans (Dietvorst, Simmons, & Massey, 2015). This preference is especially pronounced when people see algorithms err: this is when they lose confidence in automated forecasts or algorithmic judgments more quickly compared to when they see a human make a mistake (Dietvorst et al., 2015; for a review see Burton, Stein, & Jensen, 2020).

And yet other frameworks suggest that people should prefer AI and its decisions. The Modality–Agency–Interactivity–Navigability (MAIN) model draws on the work on peripheral/heuristic versus central/systematic information processing to advance the notion of “machine heuristic,” or a mental shortcut triggered when users believe they are dealing with a machine (Sundar, 2008). “When the perceived locus of our interaction is a machine, rather than another human being, . . . we automatically apply common stereotypes about machines, namely that they are mechanical, objective, ideologically unbiased and so on” (Sundar & Kim, 2019, p. 2). Because individuals tend to see algorithmic judgment as superior to human judgment in accuracy and predictive power (Dawes, Faust, & Meehl, 1989) and algorithmic outcomes as objective, reliable, and ideologically blind (Dijkstra, Liebrand, & Timminga, 1998), people evaluate AI and its decisions more favorably than parallel decisions of a human. Studies within the MAIN framework support the machine heuristic (Sundar & Kim, 2019) and the work on algorithm appreciation also shows that people are more likely to follow advice from an algorithm than a person for various tasks (e.g., forecasting, suggesting songs; Logg, Minson, & Moore, 2019) and that AI decisions are evaluated as at least as useful or fair as those of humans across
domains (media, health, and judicial) and low- or high-stakes consequences of the decisions (Araujo, Helberger, Kruikemeier, & de Vreese, 2020; see also Thurman et al., 2019).

**Users’ reactions to AI in political contexts**
It is unclear whether these findings apply to political contexts because most work comes from such rather uninvolving scenarios as hiring or consumer decisions or health recommendations. The research on users’ reactions to political moderation is relevant, finding that users attribute bias for content deletion to other humans rather than algorithms (Myers West, 2018; Shen et al., 2018), prefer combined algorithmic and social curation of harassment (Jhaver, Ghoshal, Bruckman, & Gilbert, 2018), and accept online moderation systems that are transparent (Brunk, Mattern, & Riehle, 2019). This work does not compare users’ reactions to various agents, however. Here, studies on automated journalism are informative, as they explore reactions to news attributed to journalists versus AI. The findings are mixed: sometimes news attributed to journalists is seen as more credible, perhaps because journalists activate authority heuristic (Graefe 2020; Waddell, 2018), sometimes machine-written news is seen as more objective, likely due to assumed objectivity (Liu & Wei 2019; Wu 2020), and a meta-analysis shows that ratings depend on the outcome and the method (Graefe & Bohlken, 2020). This work tests a single context and not political news, underscoring the need to systematically study users’ reactions to humans, AI, and AI-assisted humans across political situations online.

We argue that machine heuristic is most applicable to such situations, and so we should see an “AI advantage.” Humans are known to have various biases, belong to political groups, and be influenced by emotions (Marcus, Neuman, & MacKuen, 2000; Taber & Lodge, 2006), and so human political content moderation, generation, or recommendation may be seen as self-interested or persuasive in intent. For instance, a person could attribute a human moderator’s deletion of an uncivil post to the biases of the moderator, for example, their disagreement with the post (Myers West, 2018; Shen et al., 2018). In contrast, in accordance with the machine heuristic, algorithms are believed to lack affect and emotions, but not cognitive abilities, such as logic and rationality. Also, they are trained on large datasets to follow the same procedures every time, and thus are perceived as not being influenced by emotional factors (see Castelo et al., 2019a). The fact that people “are willing to grant machines more cognitive than emotional abilities” (Castelo et al., 2019a, p. 24), should give AI an advantage over humans in political contexts. In the example above, the deletion would be attributed to the fact that the post indeed contained incivility, which the algorithm successfully detected. It follows that people should evaluate AI and its political decisions more positively than a human and their decisions.

It needs to be noted that although the assumption of neutrality is reflected in public discourse, where AI is presented as fair and objective (Gillespie 2014), there is growing work arguing the contrary. AI-based tools are a black box, frequently created for purposes that are not neutral (Gorwa, Binns, & Katzenbach, 2020; Kitchin 2017), and often accused of implicit bias. The training datasets used to mimic human decisions may contain biases rooted in culture and history and relying on unrepresentative data or the presence of biases in human developers may introduce further bias into AI systems (Broussard, 2018; Noble, 2018). In fact, researchers find racial bias in judicial algorithms (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018) and show that machine learning, used to derive AI by discovering patterns in data, replicates human racial or gender biases (Caliskan, Bryson, & Narayanan, 2017; Crawford et al., 2019).

Yet, “these elaborated notions of machine weaknesses may not be at the forefront of a user’s mind . . . . Instead, the most immediately accessible ideas about machines are the ones that are well-entrenched in our mental models, which tend to be the positive stereotypes about machine infallibility
and neutrality” (Sundar & Kim, 2019). This machine heuristic may generate an AI advantage in political settings online.

Similarly, we suspect that AI-assisted humans and their decisions will be evaluated more positively than those of a human alone. The reviewed literature suggests that the mix of algorithmic precision and the “human touch” could be particularly strong by making AI more transparent and easier to understand (Dietvorst, Simmons, & Massey, 2018; Lee, 2018; Yeomans, Shah, Mullainathan, & Kleinberg, 2019). Thus, communicating to users that a decision that harnesses human and AI agency should improve perceptions of the agent. Scant evidence tests people’s perceptions of AI-assisted humans, especially vis-a-vis AI and humans. The work on automated journalism compares the three agents as news authors, sometimes showing no differences in perceived credibility of articles attributed to a journalist, machine, or both agents (Tandoc, Yao, & Wu, 2020; Wolker & Powell, 2018), and sometimes finding that the combined authorship is rated more highly and seen as less biased than AI authorship (Waddell, 2019a, 2019b).

The work reviewed above highlights the diversity of situations in which AI performance is assessed, often with metrics tailored to the context at hand. Because our project tackles multiple tasks, it necessitates a framework that is applicable across scenarios and to political settings, in which these scenarios occur. The work on organizational justice, which explores how organizations delegate tasks to algorithms with the specific goal of contrasting human and algorithmic performance and decision-making (Lee, 2018; Otting & Maier, 2018), offers such an overarching framework. This literature assesses the evaluation of decision-making using constructs tapping fairness, trust, and legitimacy. Perceived fairness is a key aspect that describes how fair people see decisions involving them to be (Colquitt, 2001; Lee, 2018; van Dijke, De Cremer, & Mayer, 2010), one that relates to trust (i.e., is the agent trusted to perform the action) and perceived legitimacy (i.e., does the agent have the right to perform an action or decision; Otting & Maier, 2018; van der Toorn, Tyler, & Jost, 2011). We also examine perceived bias, a construct particularly relevant to political contexts.

We expect that individuals will perceive AI and an AI-assisted human and their decisions as more just than a human and its content moderation, generation, and recommendation decisions (H1). Given the limited evidence comparing humans, AI, and AI-assisted humans, we explore whether an AI-assisted human or AI and its decisions will be seen as more just (RQ1)?

**AI and biased information processing**

In addition to examining users’ reactions to content decisions (Experiment 1), we test their openness to cross-cutting political information attributed to the three agents (Experiment 2). Theories of confirmation bias and motivated reasoning (e.g., Taber & Lodge, 2006) establish that people do not objectively react to content that challenges their priors, but process it in a biased way, reject it as low quality, or dismiss its source as biased. On average, therefore, people should be less open to cross-cutting content and its source than to congenial content (H2).

Yet, extending our core argument, we predict that people will be more open to such content if it is attributed to AI than to a human. As aforementioned, political information in general, and especially when counter-attitudinal, could be seen as persuasive in nature when coming from a person. Persuasion theories show that when people think that someone aims to persuade them, they process messages differently than when they do not infer persuasive intent: they are more critical, counter-argue, and see the source as untrustworthy (Campbell & Kirmani, 2000). Because, AI is not likely to be seen as having political self-interest, it should not trigger these defensive processes, per machine heuristic (Sundar, 2008). People’s reactions to the source may “spill over” to their reactions to the content.
itself, inasmuch as the receiver’s assessment of the source influences the acceptance of the message (McGuire, 1985).

Some evidence indeed shows that people agree more with the same argument when it comes from an expert system (AI) than from a human, suggesting that user acceptance of an argument from AI is not based on a thorough examination of the argument but on the heuristic of its superiority (Dijkstra et al., 1998). Other studies similarly show that news from partisan outlets is rated as less biased when attributed to AI and an AI-assisted journalist than to a human alone (Waddell, 2019b). In the explicitly counter-attitudinal context tested, we may not see the relative superiority of an AI-assisted human over a human, because the former could be seen as using AI in a biased way when presenting cross-cutting content.

We predict that individuals who are told that information that counters their prior attitudes was generated by AI will be more open to the information than their counterparts informed that the information comes from a human (H3a) or from an AI-assisted human (H3b). In testing these predictions, we focus on outcomes central to the information processing literature (Taber & Lodge, 2006) and to persuasion theories more broadly (McGuire, 1985): credibility and political bias of a source, and also agreement with and perceived quality of the message itself.

To establish that AI can in fact open people to cross-cutting information, we propose another ambitious test. Political psychology and political communication (e.g., Taber & Lodge, 2013; Wojcieszak, Azrout, & deVreese, 2018) establish that the strength of individual priors influences how people evaluate information in general and counter-attitudinal information in particular, and so the negative reaction to cross-cutting content should be especially pronounced for people with stronger political predispositions (H4). Yet, and speaking to the suggested advantages of AI, this negative reaction should be weakest when the source is AI and strongest when it is a human, such that the difference between people’s openness to counter-attitudinal information coming from AI versus a human will be larger among those with stronger political predispositions compared to those with weaker priors (H5). We do not expect such differences for congenial content.

Methods

We tested our preregistered hypotheses in two experiments, each in the United States, Spain, and Poland. To ensure that the effects do not depend on any particular context, we selected countries that differ in party system (two-party, moderate pluralism, fragmented pluralism), institutional system (presidential, parliamentary monarchy, unitary semi-presidential), and news trust (50% United States; 40% ES, 69% PL; Newman, Fletcher, Schulz, Simge, & Nielsen, 2020). Germane here, the United States and Spain have different levels of AI adoption (McKinsey & Company, 2018), with Poland having lower levels of investments in AI (McKinsey & Company, 2017). The countries also differ in their favorability toward algorithmic news recommendations (PL most, ES moderate, US least, Newman et al., 2016) and toward robots (PL more positive than ES; Gnambs & Appel, 2019). Data collection and exclusion strategies, measures, and analyses were preregistered at the Open Science Framework (https://osf.io/spxmn; the deviations are noted throughout and the preregistered analyses are detailed in Online Appendices). The replication data are available at Harvard Datavarse https://doi.org/10.7910/DVN/U0RGGY and all replication study materials are available in online appendices.

In the United States and Spain, sampling and recruitment were done by Dynata, which maintains a large panel of American/Spanish adults recruited via verified and certified sources, subject to multiple layers of authentication, and invited to take part in studies. In Poland, we used a quality panel of 286,000 participants maintained by Panel Ariadna. To enhance the panel’s quality, the company invites
citizens to join and sends rewards to its panelists by courier to assure there are no bots and that an individual does not register multiple times. Across the countries, quotas on age, education, and gender were enforced (and on ethnicity in the United States and the autonomous region in Spain).

Figure 1 illustrates the design. At the pretest, participants reported how often they engaged in four activities: read or view content on social media; share, post, or comment on social media; comment on news sites, or news social media pages; and read comments on news sites, news apps, or news social media pages. As preregistered, those who reported never engaging in any of these activities and those who failed the attention check were automatically filtered out, and we also excluded participants who completed the study in under 48% of the median time, a standard recommended by Dynata. The final samples include 2,249 subjects in the United States, 2,301 in Spain, and 2,178 in Poland. Sample sociodemographics are presented in Online Appendix 1.

The pretest measured sociodemographics and a range of political predispositions as moderators (see Online Appendix 2 for question wording in English): ideology, political identity strength (Huddy, Mason, & Aarøe, 2015), and partisanship in the United States and government support/opposition in PL. The pretest also assessed attitudes toward three issues, using three pairs of opposing statements per issue: one indicating a liberal position (e.g., "Immigrants strengthen our country because of their hard work and talents") and the other a conservative position (e.g., "Immigrants today are a burden on our country because they take our jobs, housing, and healthcare"), with a slider allowing people to place themselves in between each pair. We selected distinct issues that are salient in each country to ascertain that the effects are not attributable to any specific issue: immigration and social security in the United States, immigration and Catalan independence in Spain, and gender equality and healthcare in Poland. In addition, we selected an issue related to forecasting in the context of economic policies of Trump (United States) or the government (Spain, Poland). This variation minimizes case-category confounding, with no a priori predictions about issue differences.

Experiment 1: After the pretest, people were randomly assigned to one of the agent conditions: AI, human, or AI-assisted human. Depending on the condition, they read a description of a human

---

**Figure 1** The flowchart of the overall design.

Notes: The figure presents sample and cell sizes for the second dataset for PL (used for Experiment 2). The sample details for Dataset 1 in PL are: Total finished = 2,646, Excluded for: failed screening n = 33; failed attention check n = 99; speeders = 338; final participants = 2,178.
moderator, AI, or an AI-assisted human (68–75 words; see Online Appendix 3) and saw four scenarios in which AI is or can be used in political contexts: (a) [Agent] deletes hate speech and uncivil political posts and comments, (b) [Agent] reminds users to be civil and warns users that their behavior is inappropriate, with the goal of encouraging civility, (c) [Agent] presents political information and facts pulled from diverse media outlets to improve online discussion quality (i.e., neutrality, accuracy, factual nature), and (c) AI/editors or journalists/editors and journalists aided by AI make decisions about what news to show and recommend to readers. See Online Appendix 4 for each scenario.

After each scenario, participants answered four questions adjusted from the organizational justice literature (Lee, 2018; Otting & Maier, 2018). We asked about their perception of the agent as fair, the decision as trustworthy and legitimate, and the agent as politically unbiased. The four items per scenario loaded on the same factor and scaled well (Cronbach’s alphas around .90; the majority of eigenvalues above 3.0), and were averaged into an aggregate perceived justice outcome (see Online Appendix 5 for item wording, Cronbach’s alphas, and factor loadings per country and scenario).

For Experiment 2, participants stayed in the agent condition and were randomly assigned to a pro- (50%) versus counter-attitudinal (50%) information condition, in a three (agent: human, AI, AI-assisted human) by two (leaning: pro-, counter-attitudinal) between-subjects design. They were told they would read statements written/generated by “a person who/AI that/a person assisted by an AI that tracks relevant political news.” Depending on the condition, they saw a pro- or counter-attitudinal statement on each issue. The statements were of parallel length (see Online Appendix 6). The assignment was based on averaged pretest attitudes per issue: if a person was below/above the midpoint (i.e., aligned more with the liberal/conservative position; e.g., the pro-/anti-immigration stance in the example above), they saw a statement opposing their stance (e.g., anti-/pro-immigration). Following each statement, participants rated their agreement with the statement, its quality, and the credibility and the political bias of the agent. Because the items per issue scaled well and loaded on one factor (most Cronbach’s alphas over .90, and 8 out of 9 Eigenvalues above 3.0; see Online Appendix 7 for wording, alphas, and Eigenvalues), they were averaged into an openness to information outcome.

Analytical Approach

We used multilevel modeling (nlme package in R) because each participant was presented with four scenarios (Experiment 1) and three issues (Experiment 2), and so each assessment of the agent is not an independent observation but is nested within a participant. A participant ID was entered as a Level 2 variable with random intercepts in the model, to account for the nested structure of the data. Random intercept models yielded better fit than the baseline model (all p < .001), supporting our choice of MLM. The models predicted perceived justice (Experiment 1) and openness to information (Experiment 2) from the agent treatment. The scenarios/issues were entered as dummy variables to test for potential, nonhypothesized, differences between the scenarios/issues. We also interacted the agent condition with the scenario/issue to test whether the effects depended on the scenario/issue. The conditional R2s are 69.9% for the United States, 55.4% for ES, and 66.1% for PL in Experiment 1, and 68.7%, 56.1%, and 55.3%, respectively, for Experiment 2. Because a large percentage of explained variance is accounted for by the Level 2 variable, the Akaike Information Criterion shown in Online Appendix Tables better reflects model performance. In the text, for ease of interpretation, we present figures plotting estimated marginal means and outline the means under the figures. Online Appendix 8 for Experiment 1 and Online Appendix 9 for Experiment 2 detail model coefficients and the means

M. Wojcieszak et al. AI in Political Contexts Online

Journal of Computer-Mediated Communication 00 (2021) 1–20
with confidence intervals. Because the preanalysis plan detailed individual items as outcomes, these models are shown in Online Appendices 10 (Experiment 1) and 11 (Experiment 2).

Results

Experiment 1: Content moderation, generation, and recommendation

If our core theoretical hypotheses are correct, we would see greater perceived justice in the AI and AI-assisted human conditions than in the human agent condition (H1). Taking the human agent as the reference category, we find significant interaction effects between the AI agent and the scenarios in all three countries, suggesting that the differences between agents may be task-dependent ($p < .05$). To interpret these interactions, we plot the estimated marginal means per agent/scenario with 95% confidence intervals (see Online Appendix 8 for model coefficients and marginal means). The results, depicted in Figure 2, strongly counter our predictions. Across the countries, human agents are seen as significantly more just than AI when deleting problematic content and reminding users to be civil, and—in the United States and Poland—when presenting political facts in discussions (this difference is insignificant in Spain, yet the means suggest more positive evaluation of a human agent). The significant differences between human and AI agents average 0.50 in the United States, 0.48 in Spain, and 0.43 in Poland (7-point scale), nontrivial differences that may shift one’s reaction from neutral to positive. The deviations from this “human advantage” emerge for news recommendations: in the

Figure 2. Results of Experiment 1 per country and scenario.

Notes: The first panel compares AI to a human agent; the second panel compares AI-assisted human (AIAH) to a human agent. Marginal Means: Encouraging civil behavior. The United States: $M_{AI} = 3.92; M_{AIAH} = 4.20; M_{Human} = 4.54$; ES: $M_{AI} = 4.87; M_{AIAH} = 4.88; M_{Human} = 5.49$; PL $M_{AI} = 4.40; M_{AIAH} = 34.64; M_{Human} = 4.95$; Deleleting hate speech and incivility: The United States $M_{AI} = 3.77; M_{AIAH} = 4.01; M_{Human} = 4.26$; ES: $M_{AI} = 4.17; M_{AIAH} = 4.18; M_{Human} = 4.51$; PL $M_{AI} = 4.48; M_{AIAH} = 4.40; M_{Human} = 4.09$; Presenting diverse information: The United States $M_{AI} = 3.54; M_{AIAH} = 3.82; M_{Human} = 3.92$; ES: $M_{AI} = 3.98; M_{AIAH} = 3.83$; Recommending news: The United States $M_{AI} = 3.45; M_{AIAH} = 3.64; M_{Human} = 3.56$; ES: $M_{AI} = 3.96; M_{AIAH} = 3.82; M_{Human} = 3.64$; PL $M_{AI} = 3.52; M_{AIAH} = 3.83; M_{Human} = 3.58$. [10]
United States and Poland, there are no significant differences in how participants perceive AI versus journalists and editors, and in Spain, AI is evaluated as more just \((M = 3.93, CI = [3.86, 4.06])\) than journalists and editors \((M = 3.64, CI = [3.54, 3.73])\).

The hypothesis regarding the advantage of an AI-assisted human is similarly not supported \((H2)\), although coefficients show significant interaction effects for the AI-assisted human condition \((p < .05, \text{see Online Appendix 10})\). Participants told that the agent was a human see the decision-maker deleting problematic content and reminding users to be civil as more just than those told the agent was an AI-assisted human \((\text{the difference for the last scenario missed significance in PL})\). There are no differences between the two agents in the context of presenting facts in discussions. Only in the news recommendation scenario, in Spain and Poland, AI-assisted editors and journalists \((\text{ES: } M = 3.82, CI = [3.73, 3.92]; \text{PL: } M = 3.83, CI = [3.73, 3.93])\) fare better than editors and journalists alone \((\text{ES: } M = 3.64, CI = [3.54, 3.73]; \text{PL: } M = 3.58, CI = [3.48, 3.68])\), whereas Americans do not differ in their evaluation. These patterns lead us to reject \(H1\) and \(H2\): AI and AI-assisted humans are not seen as more just than human agents. In fact, humans have the advantage over AI in most scenarios.

We also tested for potential differences between AI and AI-assisted humans. The predicted marginal means, depicted in Figure 2, find a relatively consistent pattern in the United States and Poland. Participants who were told an AI-assisted human is deleting problematic content, reminding users to be civil, presenting facts in discussions, and making news recommendations perceived the agent and its decisions as more just than those told that it was AI making the same decisions. The average of these mean differences was less pronounced than above \((\text{the United States 0.25; PL 0.32})\). In Spain, a marginally significant effect emerged for presenting facts in online discussions, yet AI was seen as marginally more just in this case Addressing RQ1, we conclude that—for the most part—the AI-human tandem is evaluated more positively than AI alone across the various contexts in two of the three countries.

**Experiment 2: Openness to counter-attitudinal statements**

Before examining the effects of an agent, we find that across all the countries and issues, people were significantly less open to counter-attitudinal statements \((\text{the United States: } M = 2.96, CI = [2.90, 3.03]; \text{ES: } M = 2.88, CI = [2.81, 2.94]; \text{PL: } M = 3.28, CI = [3.20, 3.35])\) than to pro-attitudinal ones \((\text{the United States: } M = 4.63, CI = [4.56, 4.70]; \text{ES: } M = 4.33, CI = [4.36, 4.49]; \text{PL: } M = 4.57, CI = [4.49, 4.64]; 7\text{-point scale})\). This baseline finding, which supports \(H2\), is not surprising. The core question is whether the source of the counter-attitudinal statements makes a difference in people’s reaction. We zoom in on the participants exposed to counter-attitudinal content, with AI as the reference category.

Figure 3 plots the marginal means per issue and condition \((\text{see Online Appendix 11 for coefficients and means})\). In each country, participants are more open to counter-attitudinal information from AI than from a human on one issue: forecasts related to Trump’s economic policies in the United States (top panel), immigration in Spain (middle panel), and healthcare in Poland (bottom panel). Dissimilar content is evaluated equally negatively regardless of the source for the remaining issues: immigration and social security in the United States, Catalan independence and economic forecasts in Spain, and women’s rights and economic forecasts in Poland. The second vertical panel compares AI and an AI-assisted human. The consistent null pattern indicates no significant differences between AI and an AI-assisted human in terms of opening people up to counter-attitudinal information. These patterns do not support \(H3a\) and \(H3b\) regarding AI generating greatest openness to dissimilar political content.

The most stringent test of the potential of AI to open people up to diverse views is to examine its effects among individuals with strong political priors. We turn to heterogeneous effects, for
parsimony, describing the results for political identity strength (see Online Appendix 10 for the strength of ideology and issue attitudes, and also of partisanship in the United States and government support/opposition in PL). Those with strong political identity, on average, were indeed less open to counter-attitudinal statements than those with weaker identity (H4 supported). Although people with strong political priors were expected to be less likely to view AI offering dissimilar information as having biased persuasive intent than they would a fellow citizen offering the same information, we do not find significant differences. Figure 4 shows that the American, Spanish, and Polish participants with stronger political identities were no more open to counter-attitudinal information provided by AI than by a human on any of the issues tested. Also, as shown in Online Appendix 10, across issues and countries, we observe no significant difference between those with strong and weak priors. In short, we do not find any evidence supporting our expectation that the positive difference in favor of AI (rather than a human) presenting dissimilar information is more pronounced among those with stronger political identities. We reject H5.

Discussion

Algorithms are increasingly and seamlessly integrated into the everyday lives of all internet users. Often, social media platforms rely on AI tools to moderate content or shape users’ behaviors. Yet, we
still know relatively little about user’s perception of these tools, especially in political context that is of growing importance given the rising incivility, polarization, and misinformation spread.

Despite the expectation that AI will have a unique advantage in online political contexts, this project offered a robust finding that humans are seen as more just than AI and (often) more than an AI-assisted human as content moderators and generators. This “human appreciation,” to paraphrase Logg et al. (2019), emerged in such diverse scenarios as deleting problematic content, nudging users toward civility, and also posting diverse information in discussions. This pattern, largely consistent across three countries, suggests that human content moderation and generation is still the preferred technique to shape user behavior and contribute information. Despite the fact that this technique faces the challenges of scalability and fatigue, it is seen as more just (i.e., more credible, unbiased, trustworthy, and legitimate) than the increasingly advanced AI and also—in some cases—that AI-assisted humans. Although this appreciation of human moderators does not align with past observational work, which finds that users express “over-blown accusations of unfair moderation” (Shen et al., 2018, p. 357; Myers West, 2018), comparing various agents side by side (as done here but not in past work on moderation) may shift users’ preferences and make human moderators seem more just.

**Figure 4.** Openness to counter-attitudinal information by agent and political identity strength.

*Notes:* The first panel compares AI to a human agent as the source of counter-attitudinal information; the second panel compares AI to an AI-assisted human (AIAH) as the source of counter-attitudinal information. Estimated Marginal Means: Issue 1: Strong identity: The United States: $M_{AI} = 2.99; M_{AIAH} = 3.00, M_{Human} = 2.74; ES: M_{AI} = 2.95; M_{AIAH} = 2.97; M_{Human} = 2.70; PL M_{AI} = 4.28; M_{AIAH} = 4.26; M_{Human} = 4.29; Weak identity: The United States: $M_{AI} = 3.45; M_{AIAH} = 3.47; M_{Human} = 3.2; ES: M_{AI} = 3.39, M_{AIAH} = 3.43; M_{Human} = 3.16; PL M_{AI} = 4.07; M_{AIAH} = 3.90; M_{Human} = 3.94. Issue 2: Strong identity: The United States: $M_{AI} = 2.89; M_{AIAH} = 3.15, M_{Human} = 2.77; ES: M_{AI} = 2.85; M_{AIAH} = 3.12; M_{Human} = 2.73; PL M_{AI} = 4.99; M_{AIAH} = 4.95; M_{Human} = 5.53; Weak identity: The United States: $M_{AI} = 3.23; M_{AIAH} = 3.14, M_{Human} = 3.48; ES: M_{AI} = 3.17; M_{AIAH} = 3.10, M_{Human} = 3.44; PL M_{AI} = 4.55; M_{AIAH} = 4.64; M_{Human} = 4.33. Issue 3: Strong identity: The United States: $M_{AI} = 2.85; M_{AIAH} = 3.11, M_{Human} = 2.72; ES: M_{AI} = 2.81; M_{AIAH} = 3.07; M_{Human} = 2.68; PL M_{AI} = 4.95; M_{AIAH} = 4.78; M_{Human} = 5.06; Weak identity: The United States: $M_{AI} = 3.02; M_{AIAH} = 2.72, M_{Human} = 2.89; ES: M_{AI} = 2.97; M_{AIAH} = 2.68; M_{Human} = 2.86; PL M_{AI} = 4.13; M_{AIAH} = 4.08; M_{Human} = 4.05.
than other alternatives, especially if some users are aware of potential bias in algorithmic systems. Furthermore, an AI-assisted human was seen as most just across most situations, suggesting that users are most skeptical of AI as a content moderator and generator.

This pattern was less pronounced for recommending news. There were no differences between AI and journalists in the United States and Poland, and AI’s recommendations were seen as more just in Spain. Also, it was largely the combination of AI and journalists/editors that was evaluated most positively. This finding aligns with evidence that users see algorithmic selection as a better way to get news than editorial curation (Thurman et al., 2019). We speculate that this divergence from the pattern of “human appreciation” is due to the familiarity with the news recommendation scenario (see Castelo et al., 2019, b for familiarity). Although many people are not aware that algorithms recommend news (Fletcher, 2018), people’s mere familiarity with the scenario is likely greater than their familiarity with AI deleting problematic content, nudging discussants to be civil, and offering content in discussions. The last two scenarios are not yet widely implemented, although agents performing these roles are being developed. This likely limited experience with AI in our less common scenarios could have made AI seem more “scary” than a human. Alternatively, these results may be due to the definition of AI provided. AI was presented as a technology capable of performing the various tasks, rather than a class of technologies, with different programs carrying out defined tasks. This may have affected how people in the AI condition perceived AI, making it seem omnipotent and, again, more “scary.”

These findings and speculations underscore the continually developing sociotechnological landscape, which scholars aim to capture. To the extent that trust in algorithms increases with task familiarity and past experience (e.g., people appreciate algorithmic movie recommendations), the detected algorithm aversion may decrease as more users become aware of the rapid progress in the field of AI and of the various ways in which AI is (and will soon be) used. This could suggest that people’s reactions to AI are shaped not only (or not even primarily) by its perceived objectivity and neutrality, as posited by the machine heuristic (Sundar, 2008), but also by their familiarity or experience with AI (among other factors, e.g., how AI is portrayed in popular fiction, Sundar et al., 2016). Inasmuch as this is the case, we may see greater algorithmic appreciation moving forward. It is also possible, however, that as various instances of algorithmic failures and AI bias are revealed (Crawford et al., 2019), people may be increasingly skeptical of the role of AI in content moderation, generation, and recommendation. This possibility, which can be tracked by longitudinal work, may lead to an adaptation of the machine heuristics, in that this growing awareness could trump the perceptions of AI as ideologically blind. We advance these ideas as new avenues for future theorizing and research. We also note that the algorithmic aversion in our data may have been due to the participants being aware of the underlying problems inherent in the construction and deployment of curatorial algorithms. Unfortunately, we did not include open-ended questions to assess the motivations behind the responses.

Despite our prediction that attributing counter-attitudinal political content to AI may lower people’s defenses, we find weak and inconsistent results. In each country, individuals were more open to such content when it came from AI versus a human on one issue only, yet there were no systematic patterns by issue, whether “hot” and divisive, “colder” and complex, or one related to forecasting. We conclude that people tend to reject information challenging their views, agreeing with it less, seeing it as of lower quality, and its source as less credible and more biased, regardless of whether AI, a human, or the AI-human combination offers this information. This does not depend on the strength of individual political priors. In short, AI does not perform better than humans and AI-assisted humans as a source of cross-cutting information and—as such—its potential to minimize biased processing or generate greater understanding of dissimilar perspectives is limited. Again, these results do not align with the machine heuristic, potentially revealing its boundary conditions and the need to apply this
framework to ego-involving matters that touch on individual political attitudes. Inasmuch as partisanship and ideology are increasingly aligned with other core social identities (Mason, 2018), the possibilities of reaching individuals with dissimilar views may be limited.

Despite some limitations (e.g., lack of probability samples in the United States and Spain or the lack of open-ended questions to contextualize people’s responses), our findings have concrete theoretical and practical implications. They endorse the algorithm aversion rather than the machine heuristic (Sundar, 2008) and algorithm appreciation (Logg et al., 2019) frameworks. They may also suggest familiarity with AI and the tasks it can perform as an important factor influencing people’s reactions to AI as an agent moderating, generating, and recommending content. Research should systematically test familiarity and perception of tasks performed by AI as prevalent as well as examine these effects among non-Western audiences, whose experience with AI is likely different than that of users in Western democracies.

Practically, although social media platforms and news media are using AI to delete hate speech or misinformation, and to recommend content, many observers fear that algorithms make the public more polarized, extreme, and susceptible to misinformation (Pariser, 2011; Sunstein, 2009). Naturally, our objective was not to test if this is the case, but instead to offer AI as a potential solution that, we hoped, could be accepted by users across different political situations. Yet, algorithmic interventions are not received more favorably by social media users, nor does AI minimize individual tendency to reject dissimilar political content. Even though the general public’s perception that AI—and machines in general—are ideologically blind, lack agency, and are unaffected by emotions (Dijkstra, 1999; Dijkstra et al., 1998; Sundar, 2008), this perception may not extend to political information, especially information that counters individual priors. We again emphasize that these reactions to online interventions by AI may improve as users become more familiar with AI or—conversely—worsen if the public is made aware of biases implicit in algorithmic agents.

These ongoing sociotechnological changes underscore the ever-evolving nature of the current information and communication environment and the need for more theorizing and research. We hope that this project will invite more work on the conditions in which algorithmic interventions in political contexts can be effective. It is only when we thoroughly explore the various characteristics of AI itself and the situations in which it can intervene, and also of the users reacting to AI and its decision, can we gain a deeper insight into its present and future role in ameliorating some of the problems of the online environment and—by extension—of contemporary societies.

**Funding**

The author(s) gratefully acknowledge the support of the European Research Council, “Europeans exposed to dissimilar views in the media: investigating backfire effects,” Proposal EXPO-756301 (ERC Starting Grant), Magdalena Wojcieszak, Principal Investigator. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the European Research Council.

**Supporting information**

Additional Supporting Information is available in the online version of this article.

Please note: Oxford University Press is not responsible for the content or functionality of any
supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

1. As the term AI is used here to describe the output or capability, it also encompasses some human-designed algorithms which may output a relatively simple function of the inputs (e.g., Ofcom, 2019 for this definitional approach). Similarly, considering the multiple dimensions of the term “algorithm” (Kitchin 2017), when we use the term in this paper, we are also referring to algorithms in their capacity to perform human tasks.

2. Due to a programming error, data from Experiment 2 in PL had to be discarded. We fielded both experiments on a new sample (Figure 1) and report Experiment 1 results from the original sample and Experiment 2 results from the additional data.

3. Because those technologically savvy are more positive toward AI, the pretest measured technological skills, competences, and background (see Online Appendix 2). Randomization in both experiments was successful: the conditions did not differ on age, gender, education, ideology, and experience with technology (and partisanship, race, and income in the United States). The only exception emerged for education in the United States and so we control for this covariate.

4. If a person’s attitudes were truly neutral, we used their self-reported ideology in the US and ES (0–4—liberal, 6–10—conservatives); those who were moderate, were randomly assigned to a pro- or counter-attitudinal condition. In PL, we first relied on participants’ support-opposition toward the government; if truly neutral (5), we relied on their placement on the 11-point ideology scale, and—if truly moderate—randomly assigned to pro- or counter-attitudinal condition.

Acknowledgements

The author(s) gratefully acknowledge the support of the European Research Council, “Europeans exposed to dissimilar views in the media: investigating backfire effects,” Proposal EXPO-756301 (ERC Starting Grant), Magdalena Wojcieszak, Principal Investigator. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the European Research Council.

References

Almerekhi, H., Kwak, H., & Jansen, B. (2020). Statistical modeling of harassment against Reddit moderators. Companion Proceedings of the Web Conference 2020, 122–123. doi:10.1145/3366424.3382729

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Society, 35(3), 611–623. doi:10.1007/s00146-019-00931-w

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130–1132.

Barberà, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. American Political Science Review, 13(4), 883–901. doi:10.1017/S0003055419000352

Barberà, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? Psychological Science, 26(10), 1531–1542. doi:10.1177/0956797615594620
Bentzen, N. (2018). *Foreign influence operations in the EU - Think Tank*. European Parliamentary Research Service.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. The MIT Press, Cambridge.

Brunk, J., Mattern, J., & Riehle, D. M. (2019). Effect of transparency and trust on acceptance of automatic online comment moderation systems. *2019 IEEE 21st Conference on Business Informatics (CBI)*, 01, 429–435. doi:10.1109/CBI.2019.00056

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239. doi:10.1002/bdm.2155

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. doi:10.1126/science.aal4230

Campbell, M. C., & Kirmani, A. (2000). Consumers’ use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *Journal of Consumer Research*, 27(1), 69–83. doi:10.1086/314309

Castelo, N., Bos, M. W., & Lehmann, D. (2019a). Let the machine decide: When consumers trust or distrust algorithms. *NIM Marketing Intelligence Review*, 11(2), 24–29. doi:10.2478/nimmir-2019-0012

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019b). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825. doi:10.1177/0022243719851788

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4), 658–679. doi:10.1111/jcom.12104

Coleman, M. C. (2018). Bots, social capital, and the need for civility. *Journal of Media Ethics*, 33(3), 120–132. doi:10.1080/23736992.2018.1476149

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, 86(3), 386–400. doi:10.1037/0021-9010.86.3.386

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., … Kaziunas, E. (2019). AI now 2019 report. AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.pdf

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899), 1668–1674. doi:10.1126/science.2648573

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. doi:10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155–1170.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, 18(6), 399–411. doi:10.1080/014492999118832

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3), 155–163. doi:10.1080/014492998119526

Facebook. (2020, August). *Community Standards Enforcement*. Available at https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1(1). doi:10.30658/hmc.1.5
Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies* (pp. 167–194). The MIT Press, Cambridge.

Gnambs, T., & Appel, M. (2019). Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. *Computers in Human Behavior*, 93, 53–61. doi:10.1016/j.chb.2018.11.045

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945. doi:10.1017/2053951719897945

The Guardian (2020, September 8). A robot wrote this entire article. Are you scared yet, human?: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Graefe, A., & Bohlken, N. (2020). Automated journalism: A meta-analysis of readers’ perceptions of human-written in comparison to automated news. *Media and Communication*, 8(3), 50–59. doi:10.17645/mac.v8i3.3019

Huddy, L., Mason, L., & Aarøe, L. (2015). Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review*, 109(1), 1–17. doi:10.1017/S0003055414000604

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction*, 25(2), 12:1–12:33. doi:10.1145/3185593

Kim, Y., & Kim, Y. (2019). Incivility on Facebook and political polarization: The mediating role of seeking further comments and negative emotion. *Computers in Human Behavior*, 99, 219–227. doi:10.1016/j.chb.2019.05.022

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14–29. doi:10.1080/1369118X.2016.1154087

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237–293. doi:10.1093/qje/qjx032

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 2053951718756684. doi:10.1177/2053951718756684

Liu, B., & Wei, L. (2019). Machine authorship in situ. *Digital Journalism*, 7(5), 635–657. doi:10.1080/21670811.2018.1510740

Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press, Cambridge.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. doi:10.1016/j.obhdp.2018.12.005

Marcus, G. E., Neuman, W. R., & MacKuen, M. (2000). *Affective intelligence and political judgment*. University of Chicago Press, Chicago.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago Press, Chicago.

Masullo, G. M., Riedl, M. J., & Huang, Q. E. (2020). Engagement moderation: What journalists should say to improve online discussions. *Journalism Practice*, 0(0), 1–17. doi:10.1080/17512786.2020.1808858

McGuire, W. J. (1985). Attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, 3rd ed., pp. 233–346). New York: Random House.

McKinsey & Company (2017). Rewolucja AI: Jak sztuczna inteligencja zmieni biznes w Polsce. Available at https://www.mckinsey.com/pl/our-insights/rewolucja-ai-jak-sztuczna-inteligencja-zmieni-biznes-w-polsce
McKinsey & Company (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. Available at https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.ashx

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. Information, Communication & Society, 21(7), 959–977. doi:10.1080/1369118X.2018.1444076

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366–4383. doi:10.1177/1461444818773059

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 72–78. doi:10.1145/191666.191703

Newman, N., Fletcher, R., Levy, D., & Nielsen, R. (2018). The Reuters Institute digital news report 2016. Reuters Institute for the Study of Journalism. Available at http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf

Newman, N., Fletcher, R., Schulz, A., Simge, A., & Nielsen, R. K. (2020). Reuters Digital News Report 2020. Reuters Institute. http://www.digitalnewsreport.org

Noble, S. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press, New York.

Ofcom. (2019). Use of AI in online content moderation. Available here

Ötting, S. K., & Maier, G. W. (2018). The importance of procedural justice in Human–Machine Interactions: Intelligent systems as new decision agents in organizations. Computers in Human Behavior, 89, 27–39. doi:10.1016/j.chb.2018.07.022

Pariser, E. (2011). The filter bubble: How the new personalized web is changing what we read and how we think. Penguin Books. http://rbdigital.oneclickdigital.com

Reeves, B., & Nass, C. I. (1998). The media equation: How people treat computers, television, and new media like real people and places (1. paperback ed). CSLI Publ, Cambridge.

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. Computers in Human Behavior, 107, 106262. doi:10.1016/j.chb.2020.106262

Roberts, S. T. (2020, April 8). Fewer humans are moderating Facebook content. That’s worrying. Slate Magazine.

Shen, Q., Miller, M., Jo, Y., & Rose, C. P. (2018). Perceptions of censorship and moderation bias in political debate forums. Proceedings of the Twelfth International AAAI Conference on Web and Social Media. ICWSM

Stroud, N. J., Scacco, J., & Curry, A. (2016). The presence and use of interactive features on news websites. Digital Journalism, 4(3), 339–358. doi:10.1080/21670811.2015.1042982

Sundar, S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger ?0026; & J. Flanagin (Eds.), Digital media, youth, and credibility (pp. 72–100). The MIT Press, Cambridge.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–9. doi:10.1145/3290605.3300768

Sundar, S. S., Waddell, T. F., & Jung, E. H. (2016). The Hollywood Robot Syndrome media effects on older adults’ attitudes toward robots and adoption intentions. 2016 11th ACM/IEEE
Sunstein, C. R. (2009). *Republic.com 2.0* (1. pbk. print). Princeton University Press, Princeton.

Taber, C., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769. doi:10.1111/j.1540-5907.2006.00214.x

Tandoc, E. C., Jr., Yao, L. J., & Wu, S. (2020). Man vs. machine? The impact of algorithm authorship on news credibility. *Digital Journalism*, 8(4), 548–562. doi:10.1080/21670811.2020.1762102

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, 7(4), 447–469.

van Dijke, M., De Cremer, D., & Mayer, D. M. (2010). The role of authority power in explaining procedural fairness effects. *Journal of Applied Psychology*, 95(3), 488–502. https://doi.org/10.1037/a0018921

van der Toorn, J., Tyler, T. R., & Jost, J. T. (2011). More than fair: Outcome dependence, system justification, and the perceived legitimacy of authority figures. *Journal of Experimental Social Psychology*, 47(1), 127–138. doi:10.1016/j.jesp.2010.09.003

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. doi:10.1126/science.aap9559

Waddell, T. F. (2018). A robot wrote this? How perceived machine authorship affects news credibility. *Digital Journalism*, 6(2), 236–255.

Waddell, T. F. (2019a). Attribution practices for the man-machine marriage: How perceived human intervention, automation metaphors, and byline location affect the perceived bias and credibility of purportedly automated content. *Journalism Practice*, 13(10), 1255–1272. doi:10.1080/17512786.2019.1585197

Waddell, T. F. (2019b). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers’ evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, 96(1), 82–100.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37. doi:10.2478/jagi-2019-0002

Wojcieszak, M., Azrout, R., & de Vreese, C. (2018). Waving the red ClothMedia coverage of a contentious issue triggers polarization. *Public Opinion Quarterly*, 82(1), 87–109.

Wölker, A., & Powell, T. E. (2018). Algorithms in the newsroom? News readers’ perceived credibility and selection of automated journalism. *Journalism*, 1464884918757072. doi:10.1177/1464884918757072

Woolley, S. C., & Howard, P. N. (2016). Automation, algorithms, and politics| political communication, computational propaganda, and autonomous agents—Introduction. *International Journal of Communication*, 10(0), 9.

Wu, Y. (2020). Is automated journalistic writing less biased? An experimental test of auto-written and human-written news stories. *Journalism Practice*, 14(8), 1008–1028. doi:10.1080/17512786.2019.1682940

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4), 403–414. doi:10.1002/bdm.2118

Zuckerberg, M. (2018, April 10). Transcript of Mark Zuckerberg’s Senate hearing. Retrieved from https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/