Ethics and Information Technology (2024) 26:20
https://doi.org/10.1007/s10676-024-09759-5

ORIGINAL PAPER

Socializing the political: rethinking filter bubbles and social media with Hannah Arendt

Zachary Daus[1]

Abstract

It is often claimed that social media accelerate political extremism by employing personalization algorithms that filter users into groups with homogenous beliefs. While an intuitive position, recent research has shown that social media users exhibit *self*-filtering tendencies. In this paper, I apply Hannah Arendt's theory of political judgment to hypothesize a cause for self-filtering on social media. According to Arendt, a crucial step in political judgment is the imagination of a general standpoint of distinct yet equal perspectives, against which individuals compare their own judgments in order to test their defensibility. I argue that social media inhibit this step by gamifying the pursuit of social status, which encourages users to consider the perspectives of others not for the sake of a general standpoint but for the sake of improving their social status, resulting in self-filtering. Consequently, ameliorating political extremism on social media requires not just reforming the algorithms that deliver content to users, but the interfaces on which users present their social identities.

Keywords Hannah Arendt · Social media · Filter bubble · Polarization · Democracy · Judgment

Introduction

Over the past several years, social media platforms have exhibited a tendency to radicalize their users. The first manifestations of this phenomenon occurred not in the United States, but in countries seemingly outside the orbit of big tech. In Myanmar, Facebook aggressively sought new users by subsidizing the mobile data used to access the website, making it the de facto news source for millions of people (Fisher, 2022). It was in this context that Buddhist monk Ashin Wirathu began posting increasingly extremist content that attracted hundreds of thousands of ethnic Burmese followers and encouraged acts of genocidal violence against his country's Rohingya minority (Specia & Mozur, 2017).

The radicalizing power of social media also manifested in the United States. Instead of using social media to debate substantive issues, advisors to Donald Trump exploited their social media presence to spread claims that Democratic Party members were conspiring with the American government to sabotage his candidacy (Rosenberg, 2016). Online political discourse devolved into an escalating battle between opposing political identities, with sometimes violent offline consequences (Kang & Goldman, 2016). Only after uncovering Russian 'troll farms' with the intention of polarizing the American electorate through the spread of misinformation did American legislators begin to grasp the scope of the crisis (MacFarquhar, 2018). Now that extremism on social media was threatening American democracy from without, it could no longer be ignored.

Social media radicalization is often interpreted as the consequence of a toxic combination of algorithmic personalization and belief polarization. According to the hypothesis popularized by Eli Pariser's *Filter Bubble: How the New Personalized Web is Changing What We Read and Think* (2011), personalization algorithms sort users based on the surveillance of their online activity into groups of other like-minded users, or filter bubbles. Filter bubbles in turn induce what social psychologists refer to as belief (or group) polarization (Cho et al., 2020; Dandekar et al., 2013; Mäs & Flache, 2013; Sunstein, 2017). Belief polarization is the empirically-proven tendency for groups of individuals with similar beliefs to adopt more extreme versions of their beliefs after group interaction, such as discussion (Moscovici & Zavalloni, 1969; Sunstein, 2009). In light of these

✉ Zachary Daus
zachary.daus@monash.edu

[1] School of Philosophical, Historical, and International Studies, Monash University, Melbourne, Victoria, Australia

Published online: 21 March 2024

potentially undesirable effects of personalization, philosophers have offered different approaches for their mitigation, such as diversifying content or educating users of personalization's dangers (Alfano et al., 2018; Bozdag & den Hoven, 2015; Miller & Record, 2013; Simpson, 2012).

Empirical research has however cast suspicion on the claim that personalization-induced filter bubbles are the principal culprits for online radicalization. Recent studies have indicated that people encounter *more* diverse perspectives on social media in comparison to traditional media formats and *still* exhibit a tendency to radicalize in polarizing directions (Flaxman et al., 2016). Other studies have indicated that individuals online exhibit a greater tendency to self-select into echo chambers as opposed to being sorted algorithmically into filter bubbles (Bakshy et al., 2015; Cinelli et al., 2020; Ekström et al., 2022).[1] Further studies comparing user experience with and without personalization algorithms have problematized the correlation between personalization algorithms and radicalization (Guess et al., 2023; Kelm et al., 2023). In light of such evidence, some media theorists have concluded that the negative effects of personalization-induced filter bubbles may be overstated (Andrejevic & Volcic, 2020; Bruns, 2019). I argue that Hannah Arendt, in particular her conception of political judgment, allows us to understand how social media may radicalize users in spite of the potential absence of filter bubbles.

The political theory of Arendt is well-positioned to address the phenomenon of online extremism. A recurring theme throughout Arendt's oeuvre is the question of how individuals become radicalized and capable of immoral acts which they would not otherwise condone. In her first major book, *The Origins of Totalitarianism* (1951/2017), Arendt describes how imperialistic models of capitalism and the emergence of the bourgeoisie as Europe's leading political class created large numbers of alienated individuals vulnerable to fascist and antisemitic ideology. While *The Origins of Totalitarianism* presents a social-historical interpretation of radicalization, later books and essays, such as "The Crisis in Culture" in the anthology *Between Past and Future* (1961/2006), *Eichmann in Jerusalem* (1963/2007), "Thinking and Moral Considerations" (1971) and her posthumously published lectures on Kant's political philosophy present a more philosophical interpretation. In these works, Arendt connects radicalization with an incapacity to exercise political judgment. Adolf Eichmann's moral failure, according to Arendt, was a result not of any radically evil intentions he may have possessed, but of his banal desire for social advancement and his unwillingness to critically judge his unprecedented political situation.

For Arendt, political judgment is a process. First, individuals form an initial judgment. Then, they compare their judgment against a plurality a of distinct yet equal perspectives, which Arendt refers to as a general standpoint. Should their initial judgment seem unpersuasive, individuals are obliged to revise it or its justifications. This process of comparison can, however, become corrupted by what Arendt refers to as the *social*. When this occurs, individuals consider the perspectives of others not for the sake of arriving at a judgment that has been refined through the consideration of a general standpoint, but for the sake of improving their social status. Consequently, individuals reduce their standpoint to the perspectives of those whose agreement would positively affect their social status. Social media encourage this tendency. By enabling users to present a social identity that receives quantified approval—in the form of likes, follower requests and other kinds of online endorsement—social media encourage users to primarily consider the perspectives of those whose approval would improve the status of their social identity, thus 'gamifying' social identity. I argue that this gamification of social identity accelerates the mechanisms of belief polarization even in the presence of countervailing perspectives, and therefore offers an explanation for online radicalization without relying on the fraught concept of personalization algorithms.

The remaining paper consists of two sections separated into three sub-sections. Because Arendt's theory of political judgment is closely related to her theory of political action, Part One provides interpretations of her theories of action and judgment. This section also clarifies Arendt's conception of the social and how she understands it as a threat to political judgment. Part Two applies her theory of political judgment to social media. I argue that the gamification of social identity results in the socialization of political judgment, which accelerates the mechanisms of belief polarization even in the presence of diverse perspectives. This section also contextualizes my Arendtian approach with recent work on belief polarization by Talisse (2019;, 2021), as well as with other critiques of social media (Heersmink, 2018; Marin, 2021; Nguyen, 2021). By arguing that social media radicalization emerges at the nexus of social identity and gamification, the Arendtian approach provides a novel critique of social media that has been overlooked. I conclude with reflections on the reformation of social media in light of these considerations.

[1] Although these terms are used variably by authors, my use of 'echo chamber' refers to self-selected groups of likeminded individuals, while 'filter bubble' refers to groups that have receive personalized content. Unlike users in filter bubbles, users in echo chambers may still encounter diverse content, but are inclined to disregard it do to conformity and partisanship (Nguyen, 2020).

The political philosophy of hannah arendt

Interpreters of Arendt often divide her political philosophy into two periods (Benhabib, 2000, pp. 173–174; d'Entrèves, 1994, p. 101; Villa, 2021, p. 308): an earlier period that focuses on political action, culminating in *The Human Condition* (1958/2018), and a later period that focuses on political judgment, culminating in her posthumously published 1971 *Lectures on Kant's Political Philosophy* (1992). A chief characteristic of Arendt's theory of political action is the potential for political action to yield events that are uniquely unprecedented yet meaningful and worthy of remembrance, thereby renewing the collective meaning of a community. Arendt's theory of political judgment consequently aims to provide an account for how humans are capable of judging uniquely unprecedented actions as meaningful and therefore worthy of remembrance. This section provides an overview of her theories of political action and judgment, as well as an account of how the latter is prone to what Arendt refers to as the threat of the 'social'. This will provide a theoretical basis for my critique of social media in the following section.

Political action

In *The Human Condition*, Arendt describes political action in terms of freedom. To act politically is to act freely (Arendt, 1958/2018, pp. 30–31). Arendt develops her understanding of freedom as political action through a phenomenological comparison of two other kinds of activity: labor and work. Labor, on the one hand, is repetitive and leaves little room for unique and unprecedented action, making it the least free of human activities. Farmers, for example, are required to follow the natural cycles of the earth's climate: spring, summer, fall and winter. Work, on the other hand, lies between the unfreedom of labor and the freedom of political action. With work, Arendt has traditional craftsmanship in mind, which produces goods that are not immediately consumed and that therefore possess a degree of permanence (1958/2018, p. 138). While the creativity of work is limited by its design of a final product, it nonetheless leaves greater room for the worker's freedom of expression (1958/2018, p. 144; Villa, 2021, p. 173).

After giving phenomenological characterizations of labor and work, Arendt turns to the activity of political action. To convey her understanding of political action, Arendt employs the concepts of natality and plurality. Arendt understands natality as the human "capacity for beginning something anew" (1958/2018, p. 9; d'Entrèves, 1994, p. 67). While Arendt alludes to biological birth as an example of the actualization of this capacity, she does not mean to suggest that natality is an inherent human capacity that is capable of being exercised at will. Instead, it is a capacity that is dependent upon being initiated into a social world of humans who recognize each other as fully equal yet fully distinct, a phenomenon that Arendt refers to as "plurality" (1958/2018, p. 175; Benhabib, 2000, p. 109; d'Entrèves, 1994, p. 70). In political contexts, the actualization of natality is dependent upon joining political groups—likened by Arendt to a "second birth" (1958/2018, p. 176)—whose members recognize each other as fully distinct yet fully equal. The condition of plurality fosters the ability to act in ways that are uniquely unprecedented by giving humans an *equal* opportunity to act, regardless of how *distinct* their actions are expected to be.

Political judgment

The purpose of political action is not merely the achievement of uniquely unprecedented action, but the achievement of action that, when remembered, contributes to the shared meaning of a community (1958/2018, p. 198). Crucial to determining the meaningfulness of political action is political judgment. In her *Lectures on Kant's Political Philosophy* (1992), Arendt develops a conception of judgment suitable to political action by creatively combining Kant's views on aesthetic judgment with his views on political judgment, beginning with his distinction between "reflective" [*reflektierend*] and "determining" [*bestimmend*] judgment (Arendt, 1992, p. 83; Kant, 1790/2007, pp. 15–16). Broadly defined, determining judgment begins with a pre-existing concept under which the object of its judgment must be subsumed, while reflective judgment begins with an object for which a concept must be found (d'Entrèves, 2006, p. 250; Wicks, 2007, p. 42). Kant describes aesthetic judgment as a paradigmatic form of reflective judgment. A beautiful rose, for example, receives its predicate of beauty not because it has certain qualities that can be subsumed under a pre-existing concept of rose or beauty. According to Kant, to judge a rose as beautiful one need not even be capable of having a concept like rose or beauty. It is no surprise that Arendt, who stresses that the value of political action lies in its unprecedented uniqueness, would see political potential in Kant's theory of reflective judgment.[2]

In addition to the distinction between reflective and determining judgment, another element of Kant's theory of judgment that Arendt emphasizes is what she understands as "sociability" [*Geselligkeit*] (Arendt, 1992, p. 10). Kant claims that although reflective judgment does not rely on any pre-existing concept to judge its object, it must nonetheless strive to arrive at a judgment that could be "universally"

[2] Whether artificial intelligence could be capable of reflective judgment as interpreted by Arendt (or Kant) is an intriguing question that is nonetheless beyond the scope of this paper.

accepted by others. In his *Critique of Judgment*, Kant characterizes our ability to arrive at universal aesthetic judgments as a consequence of our sharing universal cognitive processes that cause us to experience aesthetic pleasure, or, as he describes it, "the state of mind that presents itself in the mutual relation of the powers of representation so far as they refer to give a representation to cognition in general" (Kant, 1790/2007, p. 48). Arendt omits Kant's psychological basis for the universality of aesthetic judgment and instead bases it on the free communication of distinct yet equal perspectives. Turning to his political essay "What is Orientation in Thinking?"—in which Kant suggests that rationality requires the "freedom to communicate […] thoughts publicly" (Arendt, 1992, p. 41; Kant, 1780/1991, p. 247)—Arendt claims that reflective judgment can achieve a degree of universality only when we become aware of the judgments of others. With this awareness, we can compare our own judgment with the judgments of others and revise it should it seem indefensible to them, thus arriving at a more universally accepted judgment.

While it might seem that this interpretation of Kantian sociability would lead Arendt to emphasize the significance of communication for political judgment, she instead turns to the concept of imagination. Again borrowing liberally from Kant, Arendt interprets his understanding of "imagination" [*Einbildungskraft*] as the ability to adopt an "enlarged mentality" that enables us to compare "our judgment with the possible rather than actual judgments of others" (1992, p. 43). Similar to Kant's claim that imagination is a mode of thought that disinterestedly contemplates the "mere representation" of the object of its judgment as opposed to how the object might be used or enjoyed (Kant, 1790/2007, pp. 36–37), Arendt claims that imagining the potential judgments of others ought to be performed in a mode that is "disinterested" (1992, p. 45). For Arendt, the disinterestedness of imagination refers to the ability to consider the judgments of others not for the sake of achieving self-interested goals—such as the achievement of social status or material gain—but for the sake of arriving at a "general standpoint" of distinct yet equal perspectives (Arendt, 1992, p. 44; d'Entrèves, 2006, pp. 251–252). Without a general standpoint against which we can compare our judgment, not only do we lose the ability to reflectively judge the collective meaningfulness of unprecedented political action, but risk excluding morally relevant perspectives.

The social

Disinterested political judgment can be better understood by turning to its opposite: *socialized* political judgment. While interpreters have criticized Arendt's ambiguous use of the concept of the social—one likening it to an amorphous conceptual "blob" (Pitkin, 1998)—it can be clarified as encompassing two distinct yet related meanings. On the one hand, the concept refers to the tendency for the value of politics to be reduced to its ability to advance material interests, such as the accumulation and distribution of wealth. On the other hand, it refers to the tendency for the value of politics to be reduced to its ability to achieve social goods, such as improved status. The thrust of Arendt's critique is that, once socialized, politics is no longer a condition for the possibility of spontaneously determining and redetermining collective meaning, but instead becomes a mere instrument for the achievement of pre-determined ends, namely the acquisition and distribution of material wealth and social status, respectively (d'Entrèves, 1994, pp. 58–59).

It is possible that Arendt overstates the risk of socializing politics, particularly in her critique of political approaches that emphasize a fair distribution of material goods (Bernstein, 1986). The meaning of the social that is more pertinent for this argument, however, is the tendency for politics to become instrumentalized for the sake of social goods, namely status. When this occurs, the value of political action is no longer judged by its ability to account for a distinct yet equal plurality of perspectives, but by its ability to deliver social status. In *Eichmann in Jerusalem*, Arendt characterizes the Nazi war criminal as an example of the pernicious effects of the social on political judgment, arguing that his complicity in the Holocaust was motivated not by explicitly malicious intentions, but by a "lack of imagination" that inhibited his ability to judge his unprecedented political situation (1963/2007, p. 287). While this characterization of Eichmann's historical personality has been criticized (Arendt & Scholem, 2017, p. 204; Stangneth, 2011), it offers insight into Arendt's understanding of how the social incapacitates political judgment. Reflecting her analysis of political judgment in her *Lectures on Kant's Political Philosophy*, Arendt claims that Eichmann's instrumentalization of politics for the pursuit of social status—evinced by his praise for Adolf Hitler's radical social ascent in German society to his Israeli prosecutors (1963/2007, p. 126)—had inhibited him from comparing his own judgment against those of a plurality of distinct yet equal perspectives. Eichmann instead compared his judgments against those whose agreement would result in improved social status: his superiors in the Nazi bureaucratic apparatus.

Summarizing, Arendt's conception of political judgment liberally applies Kant's theory of judgment to her unique understanding of politics and the social. Arendt focuses in particular on three aspects of Kant's theory. The first aspect, reflectivity, refers to the idea that political judgment must be capable of judging unique political action that defies pre-existing categories. The second aspect, sociability, refers to the idea that political judgment can strive towards universality only by considering the perspectives of others and by comparing their judgments with one's own judgment so as to test its defensibility. The third aspect, imagination, refers to the idea that comparison of judgments ought to be performed in a mode of disinterested contemplation, that is to say, in an attitude that considers the perspectives of others neither for the sake of social advancement—as was the case for Eichmann—nor for the sake of material gain, but for the sake of arriving at a judgment that has been informed by a general standpoint.

Political judgment on social media

The discourse around social media radicalization has shifted in the past several years. Theorists initially drew on the notion of personalization-induced filter bubbles to account for social media radicalization (Pariser, 2011; Sunstein, 2017). This interpretation has, however, been complicated by empirical and theoretical research. Empirical research has shown that users tend to radicalize on social media even when they are presented with diverse content (Flaxman et al., 2016), or when content is presented non-algorithmically (Guess et al., 2023; Kelm et al., 2023). Theoretical research has explored how the non-algorithmic characteristics of social media—such as its facility for rapidly sharing content (Rini, 2021), or its tendency to subvert trusted epistemological heuristics (Nguyen, 2023)—hinder our ability to adhere to the accepted norms of knowledge acquisition. I argue that Arendt's theory of political judgment offers an additional theoretical approach to understanding social media radicalization without relying on the empirically fraught notion of personalization-induced filter bubbles. First, I provide an overview of the argument that personalization algorithms induce radicalization through belief polarization. I then describe how the gamification of social identity accelerates the mechanisms of belief polarization even in the presence of diverse perspectives (or in the absence of personalization-induced filter bubbles). I conclude by contextualizing the Arendtian critique, describing how it diverges from yet complements other critiques of social media.

The filter bubble hypothesis

The problem of personalization algorithms and filter bubbles was first made explicit by Eli Pariser in *The Filter Bubble: How the Personalized Web is Changing What We Read and How We Think.* Pariser describes how personalization algorithms, first developed at Xerox's prolific Palo Alto Research Centre (PARC) to filter email spam (Pariser, 2011, pp. 27–28), were quickly extended to less innocuous applications, such as online advertising and social media (Andrejevic, 2019; Zuboff, 2019). The migration of filter algorithms from email inboxes to social media feeds raised the threat of belief polarization, namely the tendency for individuals to adopt more extreme versions of their beliefs after interaction with like-minded individuals (Moscovici & Zavalloni, 1969; Sunstein, 2009).

The phenomenon of belief polarization can be explained through recourse to at least two mechanisms: the informational mechanism and the social-comparative mechanism (Talisse, 2019, pp. 110 – 12). According to the former, individuals gradually radicalize in the presence of like-minded individuals because they are more likely to encounter convincing arguments or evidence in support of their already-held beliefs, and fewer countervailing arguments or evidence (Burnstein & Vinokur, 1977). According to the latter, individuals gradually radicalize as a consequence of their desire to be accepted by those who share their views. The majority of individuals, wanting "to appear to others as neither half-hearted nor as fanatical" (Talisse, 2019, p. 112), slowly radicalize as they adjust to the expressed beliefs of the more fervent (and likely more talkative) members of the group (Lamm & Myers, 1978). Social media, by algorithmically filtering both content and users, allegedly accelerate these mechanisms, leading to the radicalizing effects of belief polarization. It is worth noting that personalization-induced belief polarization will also ostensibly lead to increased *political* polarization, that is to say, ideological distance between political groups. When individuals are sorted into like-minded groups and undergo belief polarization, the distance between them will gradually grow wider, creating a spiral of belief polarization and political polarization that is difficult to reverse.

Despite the intuitiveness of the personalization algorithm argument, recent empirical research has questioned its scope. Researchers at the Nieman Foundation have concluded from empirical evidence that social media actually *increase* the tendency for users to have incidental exposure to opposing views (Fletcher & Nielsen, 2017). Research conducted by Seth Flaxman, Sharad Goel and Justin Rao support these findings, observing that social media are "associated with an increase in an individual's exposure to material from his or her less preferred side of the political spectrum" (2016, p. 298). Nonetheless, they also found evidence of an "increase in the mean ideological distance between individuals" on social media (2016, p. 298), suggesting that while political polarization is exacerbated by social media it may not be due to personalization algorithms. These findings have been confirmed by a more recent study that compared the effects of a social media feed with chronologically-presented content and a social media feed with algorithmically-presented content (Guess et al., 2023). Although the chronological feed resulted in more diverse content, it did not result in any noticeable reduction in belief or political polarization. Findings such as these appear to confirm the claim that "the more critical filter […] exists in our heads" (Bruns, 2019, p. 10).

The socialization hypothesis

Arendt's theory of political judgment, characterized as imagined sociability, offers an explanation for online radicalization without the need to invoke personalization algorithms and the filter bubbles they are said to produce. To briefly review, sociability refers to the aspect of political judgment that entails comparing our initial judgment against a distinct yet equal plurality of perspectives, or general standpoint. Should our judgments or their justifications seem indefensible, we are obliged to revise them. Imagination refers to the disinterested aspect of political judgment. When the comparison of our judgments against the perspectives of others is colored by an undo interest in our own social status or material gain, we become prone to considering other perspectives for the sake of their benefit to us, not for the sake of improving our initial judgment. This leads us to filter the perspectives of those with whom our agreement would have little impact on either our social status or material gain, thus reducing the equal diversity of our standpoint.

How do social media contribute to the socialization of political judgment and, as a consequence, online radicalization? The aims of many social media users do not only consist in communicating and relating with others, but also in cultivating and presenting—or "grooming" (Tufekci, 2008)—an online social identity through different online features. On some platforms, such as Facebook, this presentation of identity may be performed by 'liking' pages that are then present on a user profile. On other platforms, such as X (formerly known as Twitter), social identities are presented by composing a brief profile biography, choosing to follow (and accept as followers) certain users as opposed to others, and liking the content of others. All of these acts are readily visible to others, and can thus be understood as potential acts of identity signaling. For example, whether a user chooses to follow Donald Trump or Joe Biden on social media is more frequently a sign of their political allegiances than it is a sign of their commitment to staying well-informed.

To the extent that social media such as Facebook and X encourage users to present a particular social identity, it could be claimed that they already discourage the adoption of a general standpoint. This tendency is exacerbated, however, through the gamification of our social identities. Not only do social media give users specific tools of social identity presentation, they also gamify it by framing the success of social identity presentation according to metrics such as follower counts, retweets (or shares), and likes (Nguyen, 2021). That is to say, whether the online presentation of social identity is a success can be directly measured by these metrics, for example when those with whom users socially identify 'like' their activity, thus affirming their desired identity. The gamification of social identity in turn encourages users to socialize political judgment while online. In the pursuit of greater social status as measured by follower counts, shares, and likes, users of social media no longer consider their political judgments from the general standpoint of a plurality of distinct yet equal perspectives, but from the standpoint of those whose agreement would confer improved social status, such as similarly-identifying users (or users who possess the identity they seek to cultivate, in the case of the social media parvenu).

A consequence of this gamification of online social identity is the acceleration of the informational and social-comparative mechanisms that explain belief polarization, even in the presence of diverse perspectives. With respect to the informational mechanism, by encouraging users to only consider the perspectives of those whose agreement would result in improved social status, social media encourage them to ignore information originating from others with social identities that differ from their own. That social media encourage self-selection is supported by empirical research that shows individuals have a tendency to privilege online content from news and media outlets that align with their perceived social identity (e.g., conservative or liberal) (Bakshy et al., 2015; Cinelli et al., 2020; Ekström et al., 2022). With respect to the social-comparative mechanism, by allowing the success of one's presentation of social identity to be measured by gamification features, social media encourage users to continuously update their shared beliefs so as to attract the greatest number of endorsements—whether expressed through liking, sharing or follower requests—from those who share their social identity. Furthermore, because provoking content is more likely to receive more online engagement in general (Brady et al., 2021; Valenzuela et al., 2017), users seeking social status through gamified metrics will likely be inclined to post such content, thus further exacerbating radicalization.

While in these cases the radicalizing power of social identity gamification necessitates at least some degree of interaction between similarly-identifying users, it can also encourage radicalization in the absence of such interaction. According to Talisse (2021), the radicalization of belief polarization is caused not only through information sharing or the direct affirmation of in-group belonging, but through corroboration. The corroboration view states that radicalization can occur when beliefs are only *indirectly* substantiated as being representative of a social identity, such as when a liberal is exposed to polling information that liberals are more likely to oppose genetically modified food (2021, p.

219). The pleasant experience of having their social identities affirmed through corroboration encourages individuals to seek more identity-affirming beliefs, resulting in an increased commitment to the perceived perspective of their social identity and a decreased interest in the perspectives of those who do not share their social identity. Social media would encourage this mechanism not only by placing social identities at the forefront of online interactions, but by amplifying the voices of popular users who have 'won' the game of social media. These users become 'tastemakers' for the social identities they represent and a source of belief corroboration for those who follow them.

In all of these cases, the emphasis that social media place on the cultivation and presentation of social identity through gamification features encourages radicalization. This aligns with the Arendtian hypothesis outlined in the previous section, namely that the socialization of political judgment promotes radicalization. In the pursuit of social status, individuals invariably limit their standpoint to those who share their social identity, thus accelerating the mechanisms underlying belief polarization, even in presence of diverse perspectives and in the absence of filter bubbles. First, it encourages users to overlook information offered by the perspectives of individuals who do not share their social identity and that may countervail their own judgments. Second, it encourages users to primarily adopt positions that will be accepted by those who share their perceived (or desired) social identity. Finally, the pleasant feeling of affirmation that comes through belief corroboration, whether direct or indirect, strengthens their commitment to the perceived perspective of their social identity at the cost of excluding other perspectives. The pursuit of social status on social media accelerates all of these mechanisms, making Arendt's theory of political judgment highly applicable to the problem of social media radicalization.

Contextualizing the socialization hypothesis

By locating its cause at the nexus of gamification and social identity presentation, the foregoing analysis offers a novel explanation for radicalization on social media. This analysis nonetheless bears certain similarities to existing critiques of social media. Before concluding this section, I will contextualize the foregoing Arendtian analysis with these critiques. First, I turn to a critique of social media gamification that, while critical of its effects on communication, overlooks its full potential to radicalize users (Nguyen, 2021). I then turn to a virtue-epistemological critique of the internet (Heersmink, 2018). While bearing certain similarities to the vices identified by virtue epistemologists, I argue that the socialization of political judgment cannot be strictly interpreted as a vice, since the pursuit of social status is likely an important element of healthy human psychology. In order to frame the problem differently, I conclude by drawing on a recent critique of the sharing of misinformation on social media, which locates its source in the tendency for social media platforms to blur the normative boundaries between distinct social practices (Marin, 2021).

C. Thi Nguyen's critique of social media gamification bears a number of similarities with the foregoing socialization hypothesis (2021). Similar to the Arendtian critique that social identity gamification reduces the goals of online communication to the achievement of social status, Nguyen's argument likewise claims that gamification negatively alters the goals of communication. Using the social media platform Twitter (now known as X) as the target of his analysis, Nguyen describes how its design reduces the goals of communication to the achievement of likes, retweets, and follower counts. These metrics fundamentally transform both communication and communicators. First, it makes popularity the overarching goal of communication. The value of a post is directly measured by the number of likes or retweets it receives. Second, it changes the communicative goals of users. Metrics are less capable of measuring whether online communication has achieved goals that are less quantifiable, such as "quality of engagement, empathy, or depth of thought" (2021, p. 424), thus making these goals less appealing to users. Although Nguyen's criticism is compatible with and complementary to the foregoing Arendtian critique, it does not fully connect the gamification of social identity with the radicalization of users to the extent that the Arendtian critique does. By reducing the diversity of perspectives that users are willing to consider, the gamification of social identity radicalizes users in ways that other gamification metrics do not.

A second critique of internet usage that appears compatible with the foregoing Arendtian critique is the virtue-epistemological approach. According to Richard Heersmink, virtue epistemologists understand epistemic virtues as more fundamental to the acquisition of knowledge than justification, because epistemic virtues are the behaviors that make individuals receptive to justified belief in the first place (2018). If one is epistemically virtuous, justified belief will likely follow. Drawing on the tradition of virtue epistemology, Heersmink identifies a number of virtues that are relevant for social media use, the most applicable being intellectual humility (awareness of one's cognitive limitations or weaknesses), intellectual carefulness (avoidance of intellectual errors or mistakes), and open-mindedness (willingness to consider alternative views) (2018, pp. 3–4). Heersmink argues that these virtues are particularly important for combatting the vices that certain features of the internet promote, such as filter bubbles and misinformation. These virtues could also be applied to the Arendtian analysis, particularly intellectual humility and open-mindedness. The virtue of intellectual humility would alert users to the epistemic dangers of pursuing social status on social media, while the virtue of open-mindedness would encourage users to refrain from limiting their standpoint to the perspectives of those who share their social identity.

The Arendtian approach to online radicalization nonetheless differs from virtue-epistemological approaches in a fundamental way. Arendt's characterization of Eichmann, regardless of how accurate, was not that of an unusually *vicious* character, but that of a *banal* one. More precisely, the flaw in Eichmann's character was not the unusual degree of vice he harbored, but the degree to which he desired social status within the Nazi bureaucratic apparatus. Arendt's point is that the pursuit of social status in itself is not necessarily morally pernicious (unlike the pursuit of vice). In certain contexts, the pursuit of social status can in fact be morally righteous, such as for individuals whose social identities are historically marginalized and undervalued. In other contexts, however, such as when the social identities in question are two dominant political perspectives (e.g., conservative and liberal), then the pursuit of social status will likely result in radicalization and polarization. Similar to Lavinia Marin's diagnosis of the sharing of misinformation on social media (Marin, 2021), the problem of social media radicalization lies not in users who are prone to epistemic vice, but in a social media landscape that blurs the normative boundary between fundamentally different social practices: pluralistic political debate and social identity cultivation.

Conclusion

I have argued that the political philosophy of Arendt is a useful tool for critiquing the radicalizing effects of social media. More specifically, I have argued that certain non-algorithmic features of social media, namely the gamification of social identity, encourage users to socialize political judgment. When this occurs, social media users no longer arrive at judgments by comparing their own judgments against a plurality of equal yet distinct perspectives, but by comparing their judgments against the perspectives of those whose agreement would affirm their social identity and thereby improve their social status. The socialization of political judgment in turn accelerates the mechanisms of belief polarization— even in the absence of filter bubbles and in the presence of diverse perspectives—and offers an explanation for empirical findings that have problematized the correlation between personalization and radicalization on social media.

The novelty of the Arendtian approach lies in its locating the cause of social media radicalization at the nexus of gamification and social identity. In contrast, previous theoretical approaches have placed the blame for online radicalization at the level of algorithmic personalization (Alfano, 2018), individual epistemic practices (Heersmink, 2018), or skewed truth heuristics (Nguyen, 2023). If we are to take the Arendtian approach seriously, then we will have to rethink our approaches to constraining online radicalization. Of equal importance to diversifying online content and improving user practices may be a strategic 'de-gamification' of social media, particularly with respect to features that signal social status, such as public follower counts and endorsement metrics.

To conclude, it is worth stressing that social media are complex technologies whose impact on users cannot likely be explained with a single theory. The foregoing Arendtian approach to understanding how social media encourage online radicalization is likely not the only valid approach, and is best viewed as one of several tools that may all be necessary for fully understanding the effects of social media on users.

Acknowledgements I am grateful to Robert Sparrow and two anonymous reviewers for feedback on this article. I am also grateful to Mark Coeckelbergh, Wessel Reijers, Eero Arum and the conference participants of *Technopolitics: Charting the Unknown* at the University of Coimbra for feedback on earlier drafts and related research.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflicts interests None.

References

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. Journal of the American Philosophical Association, 4(3), 298–322.

Andrejevic, M. (2019). *Automated media* Routledge.

Andrejevic, M. & Volcic, Z. (2020). From mass to automated media: Revisiting the 'filter bubble'. In N. Witzleb, M. Paterson, & J. Richardson (Eds.), *Big data, political campaigning and the law:*

Democracy and privacy in the age of micro-targeting (pp. 17–33). Routledge. https://doi.org/10.4324/9780429288654

Arendt, H. (1971). Thinking and moral considerations. *Social Research, 38*(3), 417–446.

Arendt, H. (1992). *Lectures on Kant's political philosophy* (R. Beiner, Ed.). University of Chicago Press.

Arendt, H. (2006). *Between past and future: Eight exercises in political thought.* Penguin Books. (Original work published 1961).

Arendt, H. (2007). *Eichmann in Jerusalem: A report on the banality of evil.* Penguin Books. (Original work published 1963).

Arendt, H. (2017). *The origins of totalitarianism.* Penguin Books. (Original work published 1951).

Arendt, H. (2018). *The human condition.* University of Chicago Press. (Original work published 1958).

Arendt, H. & Scholem, G. (2017). *The correspondence of Hannah Arendt and Gershom Scholem* (A. David, Trans.). University of Chicago Press. (Original work published 2010).

Bakshy, E., Messing, S., & Adamic, L. (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Benhabib, S. (2000). *The reluctant modernism of Hannah Arendt.* Rowman and Littlefield.

Bernstein, R. (1986). Rethinking the social and the political. In *Philosophical profiles: Essays in a pragmatic mode* (pp. 238–259). University of Pennsylvania Press.

Bozdag, E. & den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics of Information Technology, 17*, 249–265. https://doi.org/10.1007/s10676-015-9380-y

Brady, W. J., McLoughlin, K., Doan, T. N., Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances, 7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Bruns, A. (2019). Filter bubble. *Internet Policy Review, 8*(4), 1–14. https://doi.org/10.14763/2019.4.1426

Burnstein, E. & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology, 13*(4), 315–332. https://doi.org/10.1016/0022-1031(77)90002-6

Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media, 64*(2), 150–172. https://doi.org/10.1080/08838151.2020.1757365

Cinelli, M., Brugnoli, E., Schmidt, A. L., Zollo, F., Quattrociocchi, W., & Scala, A. (2020). Selective exposure shapes the Facebook news diet. *PLOS One, 15*(3), e0229129. https://doi.org/10.1371/journal.pone.0229129

d'Entrèves, M. P. (1994). *The political philosophy of Hannah Arendt.* Routledge.

d'Entrèves, M. P. (2006). Arendt's theory of judgment. In D. Villa (Ed.), *The Cambridge companion to Hannah Arendt.* Cambridge University Press. https://doi.org/10.1017/CCOL0521641985.013

Dandekar, P., Goel, A., & Lee, D. (2013). Bised assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 5791–5796. https://doi.org/10.1073/pnas.1217220110

Ekström, A., Niehorster, D. C., & Olsson, E. J. (2022). Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports, 7*, 1–10. https://doi.org/10.1016/j.chbr.2022.100226

Fisher, M. (2022). *The chaos machine: The inside story of how social media rewired our minds and our world*. Quercus Editions.

Flaxman, S., Goel, S., & Rao, J. M. (2016) Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly, 80*, 298–320. https://doi.org/10.1093/poq/nfw006

Fletcher, R. & Nielsen, R. K. (2017). Using social media appears to diversify your news diet, not narrow it. *Nieman Lab*. https://www.niemanlab.org/2017/06/using-social-media-appears-to-diversify-your-news-diet-not-narrow-it/

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D. ,Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., et al. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science, 381*, 398–404. https://doi.org/10.1126/science.abp9364

Heersmink, R. (2018). A virtue epistemology of the internet: Search engines, intellectual virtues and education. *Social Epistemology, 32*(1), 1–12. https://doi.org/10.1080/02691728.2017.1383530

Kang, C. & Goldman, A. (2016, December 5). In Washington pizzeria attack, fake news brought real guns. *New York Times.*https://www.nytimes.com/2016/12/05/business/media/comet-ping-pong-pizza-shooting-fake-news-consequences.html

Kant, I. (1991). What is orientation in thinking? In H. Reiss (Ed. & Trans.), *Political writings* (pp. 237–249). Cambridge University Press. (Original work published 1780).

Kant, I. (2007). *Critique of judgment* (J. Meredith, Trans.). Oxford University Press. (Original work published 1790).

Kelm, O., Neumann, T., Behrendt, M., Brenneis, M., Gerl, K., Marschall, S., Meißner, F., Harmeling, S., Vowe, G., & Ziegele, M. (2023). How algorithmically curated online environments influence users' political polarization: Results from two experiments with panel data. *Computers in Human Behavior Reports, 12*, e100343. https://doi.org/10.1016/j.chbr.2023.100343

Lamm, H. and Myers, D. (1978). Group-induced polarization of attitudes and behavior. *Advances in Experimental Social Psychology, 11*, 145–187. https://doi.org/10.1016/S0065-2601(08)60007-6

MacFarquhar, N. (2018, February 18). Inside the Russian troll factory: Zombies and a breakneck pace. *New York Times.*https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html?smid=url-share

Marin, L. (2021). Sharing (mis)information on social networking sites. An exploration of the norms for distributing content authored by others. *Ethics and Information Technology, 23*, 363–372. https://doi.org/10.1007/s10676-021-09578-y

Mäs, M. & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLOS One, 8*(11), e74516. https://doi.org/10.1371/journal.pone.0074516

Miller, B. & Record, I. (2013). Justified belief in a digital age: On the epistemic implications of secret internet technologies. *Episteme, 10*(2), 117–134. https://doi.org/10.1017/epi.2013.11

Moscovici, S. & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*(2), 125–135. https://doi.org/10.1037/h0027568

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme, 17*(2), 141–161. https://doi.org/10.1017/epi.2018.32

Nguyen, C. T. (2021). How Twitter gamifies communication. In J. Lackey (Ed.), *Applied epistemology* (pp. 410–436). Oxford University Press. https://doi.org/10.1093/oso/9780198833659.003.0017

Nguyen, C. T. (2023). Hostile epistemology. *Social Philosophy Today, 39*, 9–32. https://doi.org/10.5840/socphiltoday2023391

Pariser, E. (2011). *The filter bubble: How the personalized web is changing what we read and how we think.* Penguin Books.

Pitkin, H. (1998). *The attack of the blob: Hannah Arendt's concept of the social.* University of Chicago Press.

Rini, R. (2021). Weaponized skepticism: An analysis of social media deception as applied political epistemology. In E. Edenberg & M. Hannon (Eds.), *Political epistemology* (pp. 31–48). Oxford University Press.

Rosenberg, M. (2016, December 5). Trump advisor has pushed Clinton conspiracy theories. *New York Times.*https://www.nytimes.com/2016/12/05/us/politics/-michael-flynn-trump-fake-news-clinton.html

Simpson, T. (2012). Evaluating Google as an epistemic tool. *Metaphilosophy, 43*(4), 426–443. https://doi.org/10.1111/j.1467-9973.2012.01759.x

Specia, M. & Mozur, P. (2017, October 27). A war of words puts Facebook at the center of Myanmar's Rohingya crisis. *New York Times.*https://www.nytimes.com/2017/10/27/world/asia/myanmar-government-facebook-rohingya.html?smid=url-share

Stangneth, B. (2014). *Eichmann before Jerusalem: The unexamined life of a mass murderer* (M. Ruth, Trans.). Knopf. (Original work published 2011).

Sunstein, C. (2009). *Going to extremes: How like minds unite and divide.* Oxford University Press.

Sunstein, C. (2017). *Republic: Divided democracy in the age of social media*. Princeton University Press.

Talisse, R. (2019). *Overdoing democracy: Why we must put politics in its place*. Oxford University Press.

Talisse, R. (2021). Problems of polarization. In E. Edenberg & M. Hannon (Eds.), *Political epistemology* (pp. 209–225). Oxford University Press.

Tufekci, Z. (2008). Grooming, gossip, Facebook and Myspace: What can we learn about these sites from those who won't assimilate? *Communication and Society, 11*(4), 544–564. https://doi.org/10.1080/13691180801999050

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral effects of framing on social media users: How conflict, economic, human interest, and morality frames drive news sharing. *Journal of Communication, 67*(5), 803–826. https://doi.org/10.1111/jcom.12325

Villa, D. (2021). *Arendt.* Routledge.

Wicks, R. (2007). *Kant on judgment.* Routledge.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York: PublicAffairs.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.