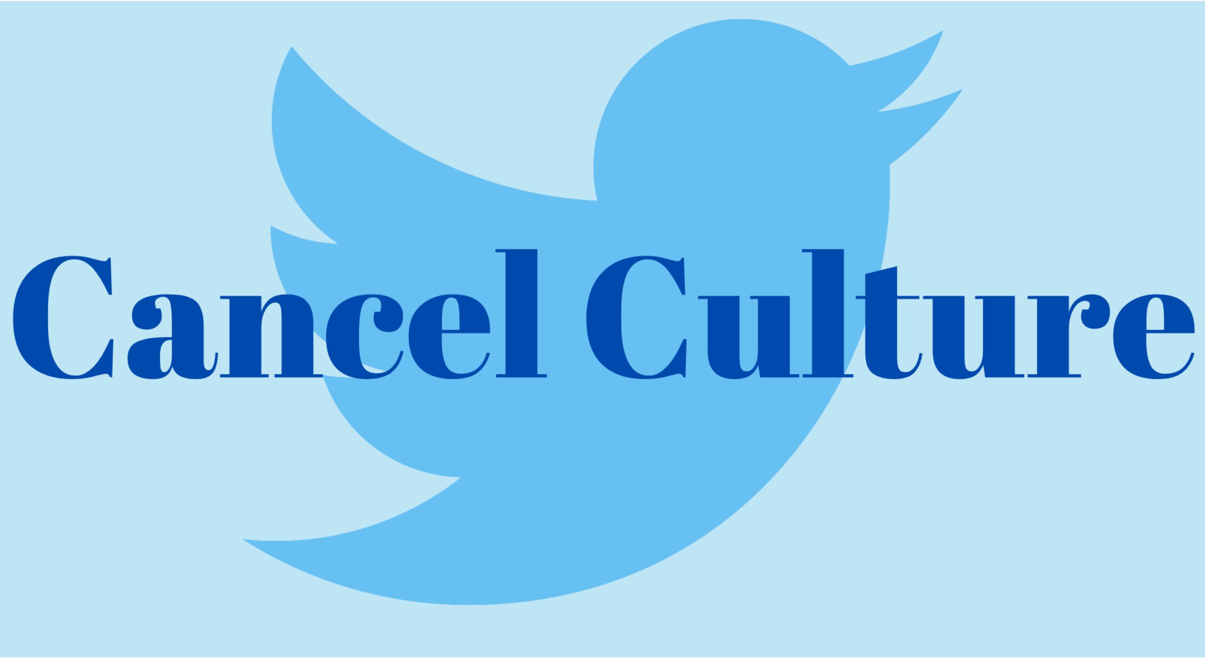
Discourse in the Age of Cancel Culture: An Analysis of Twitter's Polarising Conversations



[1]

Fabienne Plieger

6152058

Department of Interdisciplinary Social Sciences, Universiteit Utrecht

Bachelorproject ISW (201300034)

Janna de Graaf

02.07.2021

Word count: 7611

[1] https://cherwell.org/2020/04/28/public-enemy-number-one-cancel-culture-and-its-targets/

Abstract

Public prosecution is nothing new but online social networking platforms such as Twitter have taken the social phenomenon to a new level with a brand new name: *'canceling'*. As canceling is becoming a more prominent and influential social phenomenon, a so-called 'cancel culture' has started to form. Cancel culture is a term often meant as a critique by people who believe the judgements made about the canceled individual(s) have created a culture of hypercriticism. Theories on intra- and inter-group polarization are utilized to understand why discourse on Twitter about canceling a person has the ability to lead to group polarization. By collecting 717 Tweets and doing a qualitative analysis in Nvivo multiple insights were found. Social comparisons appear to have an exacerbating effect on the opinions of the users along with the group memberships of the Twitter users as well. Future research on these found effects are needed as the subject is badly under-researched, yet could possibly have a significant impact on group polarization.

Introduction

Public prosecution is nothing new but online social networking platforms such as Twitter have taken the social phenomenon to a new level with a brand new name: *'canceling'*. Merriam-Webster has recently added the extended meaning of 'to cancel' to their 'Words We're Watching' list. Their newly minted definition of 'to cancel someone' being: to stop giving support to a person (usually a celebrity or well-known figure) in response to their objectionable behavior or opinions. The act of canceling can include boycotts or a refusal to promote or engage with the person's work(s) (Editors of Merriam-Webster, 2021).

People who cancel other people have themselves been met with criticism of being hyper-judgemental. As canceling is becoming a more prominent and influential social phenomenon, a so-called 'cancel culture' has started to form. Cancel culture is a term often meant as a critique by people who believe the judgements made about the canceled individual(s) have created a culture of hypercriticism. Cancel culture can be seen as a subculture within 'woke culture': a culture in which people aim to be aware of social inequalities and strive for a better, more just society (Editors of Merriam-Webster, 2017).

Examples of people getting canceled and its impact vary to the extremes: some canceled celebrities have lost paid partnerships (Lea Michele), jobs (Hartley Sawyer), and some have thrived nonetheless (J.K. Rowling). Michele was accused of racist microagressions and racist remarks during her time working on the FOX television show 'Glee'(BusinessInsider, 2020). Sawyer got laid off from the CW television show 'The Flash' after racist and misogynist tweets resurfaced from 2012-2014 (BBC News, 2020). Rowling faced intense online backlash after writing tweets about womanhood that were seen by many as transphobic but has, as far as the public has knowledge of, not faced any complications in her career (NOS, 2020). The ramifications of being canceled don't seem to be uniform.

The cancelations are also socially ambiguous: whether someone is or is not canceled is not entirely clear. Relationships like the one between the actors Amber Heard and Johnny Depp especially prove this. Allegations from both sides with multiple trials and evidence from the cases being made public in a time of 'fake news' has created an incredibly complex web of events with heavy accusations from both sides. Their relationship has sparked fierce debates about domestic abuse from both sides with Heard and Depp both being possible victims and/or abusers. The relationship between Depp and Heard demonstrates just how polarizing online debates can become. In 2020 Depp was asked to resign his role in the Fantastic Beasts franchise after being denied libel in his trial against the Sun for calling him a wife-beater (Delbyck, 2020; Sarkisian & Ntim, 2021;). In retaliation an online petition was started to recast Heard in her upcoming movie 'Aquaman 2', gaining over 1,5 million votes (Change.org, 2020). Every day both parties get harassed online by supporters from opposite sides.

The effects of the indiscretions of celebrities and other well-known figures reach far beyond them and their careers. Each time someone is canceled it is brought on by public debate both on- and offline, and the debates are often polarizing. In this thesis group polarization will be using the following definition (with the caveat that the small group will be the study population): "Group polarization is a phenomenon in which the opinions held by members of a small group become more extreme after the group discusses a topic." (Turner & Smaldino, 2020).

Polarization is of incredible importance to understand. According to Turner and Smaldino (2018) polarization can endanger the functions of our democratic society through its destabilizing effects on clear communication amongst citizens, accurate interpretations of facts, and humanizing interactions. Discourse around cancel culture has been the subject of a lot of media attention and often heated public debate, but virtually nothing is known about it. It is therefore vital to gain a better understanding of how online discourse is becoming polarized by the subject of cancelations. Consequently this thesis's research question is: How is online discourse on Twitter about canceling (someone) leading to group polarization?

Theoretical framework

Group polarization

The aim of this study is to understand why discourse on Twitter about canceling a person has the ability to lead to group polarization. Therefore, polarization as a phenomenon first has to be further elaborated on.

Multiple theories have been offered by social psychologists to explain group polarization, two well established ones being: persuasive arguments theory and social comparison theory.

Persuasive arguments theory

Persuasive arguments theory (PAT) posits that when individuals discuss a topic with members of a biased group they gather more persuasive arguments that support their biases and lead to even more extreme ones (Bishop & Myers, 1974; Vinokur & Burstein, 1974; Turner & Smaldino, 2020).According to the PAT moderate opinions exist only because of a lack of arguments for a more extreme opinion (Turner & Smaldino, 2020). The theory also predicts a counter polarising effect, based on computational models analyzed by Mäs and Flache (2013). According to PAT, when opposing groups interact their opinions will start to become more similar. This happens because through interacting with members of an opposing group new counterarguments will be introduced, which reduces the extremism of the group towards a more moderate opinion (Turner & Smaldino, 2020).

The persuasive arguments theory requires an experiment to be proven, thus can therefore not be proven to be of influence in this thesis. However, PAT can possibly offer an explanation if neutralized conversations are found in the data, because of the mitigating effect described by the theory. Both the polarizing effect and the mitigating effect of the PAT have been proven by Mäs and Flache in 2013 and their findings were supported by Turner and Smaldino's experiment in 2020.

Social comparison theory

Social comparison theory (SCoT) posits that privately-held opinions by individuals are often more extreme than the ones voiced publicly. Exposure to like-minded people gives the individual the confidence to express their more extreme opinions openly (Myers, 1982; Turner & Smaldino, 2020). Positive social feedback such as encouraging body language or words or through another person stating a similar extreme opinion can supply the encouragement for an individual to voice their more extreme opinion. SCoT presumes that opinions are intrinsically extreme, it is the natural state of opinions. Group members calculate (primarily subconsciously) what the optimal opinion is to voice publically, taking into account their privately held opinion and the social consequences they believe they would face if they voiced that opinion publically. According to Myers's experiments in 1977, 1978, and 1980, and Smaldino's in 2017 and 2019, the post group discussions optimal public opinion is often more extreme than the one the group member initially voiced, after comparing opinions with the other group members (Myers, 1982; Turner & Smaldino, 2020).

According to SCoT this would be an expression of the true opinion the group member held all along, however Myers did question whether the social comparisons could also cause a ‘one-upping’ effect. People could want to voice a more extreme opinion to enhance their presentation of themselves towards the group (Myers, 1982). Through the manner of data collection and the data itself in this study, one reason for voicing a more extreme opinion after social comparison or another cannot be distinguished. However, social comparisons themselves will be analysed to review whether positive social feedback or similar extreme opinions stated by others lead an individual to reply with more extreme statements.

Stubborn extremism theory

Turner and Smaldino (2020) have formulated a third theory: stubborn extremism theory (SET). Stubborn extremism is not in conflict with the previously mentioned two theories, rather it explains a possible underlying mechanism of group polarization that can fit within either of the theories. Stubborn extremism describes how, when a person’s opinion on a topic becomes more extreme it also becomes more stubborn. The opinion of the individual becomes less susceptible to social influence: a person is less likely to be swayed from their opinion. Once an extreme opinion is reached they are held on to tighter (Turner & Smaldino, 2020).

Through agent based modeling Turner and Smaldino have conducted experiments in 2020 to demonstrate that stubborn extremism indeed gives rise to group polarization. Their model allowed for both positive and negative influences in which similar agents could become more similar after interactions, whilst the initially dissimilar agents could become more polarized. The expectations they had for their model were confirmed, the results were in line with both the PAT and the SCoT, but also showed that stubborn extremism occurred once an extreme opinion was reached (Turner & Smaldino, 2020).

Inter-group conflicts

The theories on group polarization help to explain how an individual’s (voiced) opinion on a subject can become more extreme through discussion with like minded people.

However, as mentioned in the previous subchapter, intergroup differences also play a role in group polarization. Understanding how groups that hold different opinions interact with each other is of vital importance to answer the research question, therefore the social identity approach (SIA) will be used to form a better understanding of the intergroup relations. SIA consists of the social identity theory (SIT) and the self-categorization theory (SCT).

Social identity theory (SIT)

The SIT argues that human interaction can be put on a spectrum ranging from purely interpersonal to purely intergroup (Hornsey, 2008; Tajfel &Turner, 1979). The purely interpersonal interactions are between people purely as individuals without any awareness of social categories. The purely intergroup interactions are between people as purely representatives of the groups they belong to. Their individual characteristics are overshadowed completely by the salience of their group membership (Hornsey, 2008; Tajfel &Turner, 1979). When intergroup interactions become more dominant, 'us versus them' distinctions can be made based on category distinctions.

Category distinctions are social categories based on markers such as ethnicity, political affiliations, cultures, etc. By making distinctions such as 'us' and 'them' an individual's concept of self becomes tied to the social identity of the group's membership. This explains why people favour the groups they are members of over the ones they are not. Tajfel and Turner (1979) believed that underlying competitive behaviour between groups is motivated by people wanting a positive and secure self-concept , meaning that the self is hereby in part formed by the group (Hornsey, 2008; Tajfel &Turner, 1979).

In order to gain or keep positive group identity multiple patterns of behaviour have been identified: derogation of the other group (the out-group)and/or glorifying one's own group (the in-group) by diminishing flaws of the in-group, engaging in social change to create a social hierarchy more in the in-group's favour and/or discriminating against the out-group (Hornsey, 2008; Tajfel &Turner, 1979). The desire to develop and maintain a positive image of one's own group, to defend it and justify actions of fellow ingroup members is called group justification (Jost, Banji & Nosek, 2004).

In contrast to the social cognitive view of stereotypes, in which they are over-simplifications as a result of a limited capacity to process social information, the SIT posits that stereotypes occur as a result of a meaning-seeking process influenced by social historical context. Stereotypes help legitimize the past and current actions of the in-group (Hornsey, 2008; Oakes, et al., 1994; Spears, et al., 1997; Tajfel, 1981).

Insights from the SIT can be applied to cancel culture: the in-group being seen as woke people and the out-group being seen as people who still support the canceled individual or are against cancel culture altogether. The woke in-group would then seek to remain or attain a positive and secure self-concept by engaging in behaviours such as glossing over flaws within their woke in-group, calling out negative behaviours or viewpoints (in their held belief) of the out-group, glorifying their own morality. Vice versa for people who are anti-cancel culture (or defenders of a canceled individual) being the in-group.

Self categorization theory (SCT)

The SCT was created after the SIT and instead of using the spectrum of social interaction it focuses on self-categorisation. SCT proposes that three levels of self-categorisation are important to the concept of self: human identity, social identity, and personal identity[2] (Hornsey, 2008; Turner et al., 1987).

In different contexts different categories will be used as the basis for an individual's identity. Which category is salient, is dependent on accessibility (whether a certain category is available in a particular context) and fit (the level to which the category is seen as accurately reflective of the real word). If social behaviour and group membership are in line with stereotypical expectations then the category distinction is more likely to have a high fit (normative fit). A high fit can also be perceived if the inter-category differences are maximized and the intra-category differences are minimized with a particular category distinction (comparative fit) (Hornsey, 2008; Oakes, 1987; Oakes, et al., 1991).

An important concept of the SCT is depersonalization. Within the context of the theory depersonalization describes how, when a category becomes apparent, people start to see themselves and other members of the same category less as individuals and more as interchangeable group prototype representatives (Hornsey, 2008; Tajfel &Turner, 1979).

If category distinctions become salient it is very possible that depersonalization could occur. Depersonalization could then explain woke people as part of a larger culture (woke culture). Individuals become representatives of woke culture and it's values. The group an individual belongs to is, according to the SIT, tied to their concept of self and self-esteem therefore, in order to keep those secure and positive the group has to have a good social

[2] Human identity is the category of the self as a human being. Social identity is the category of the self as a member of a social in-group (defined by its contrast to out-groups). Personal identity is the category of the self based on interpersonal comparisons (Hornsey, 2008; Turner et al., 1987)

standing. The individual becomes an active actor in securing the positive image of the group membership once depersonalization occurs.

On Twitter, category distinctions can become apparent through an individual's username, profile picture, handle and/or word use. This study is anonymous so only word use can be taken into account, but it is possible and likely that these might as well play a role in the interactions between people.

Groupthink

Myers stated that (1982) group interaction could be a magnifying factor in social conflicts, because in 1976 Myers and Bach's laboratory observations of people in which the results demonstrated that people are more likely to justify their own behaviours when in conflict when they were part of a group than when they were alone. Myers attributes this group-enhanced self-justification to Janis's 'groupthink' theory from 1972. Janis describes groupthink as "a psychological drive for consensus at any cost that suppresses dissent and appraisal of alternatives in cohesive decision making groups" (1972). Janis states that thoughts deviating from the group's norm are suppressed in an effort to seek concurrence, mutual support in order to maintain self esteem (Janis, 1972; Myers & Lamm, 1975).

Janis's groupthink has a strikingly similar explanation for group-enhanced self-justification as the SIT has for group justifications, as both attribute the justifications to an individual's desire to maintain or seek positive self-esteem. Janis's groupthink differs from group justification of the SIT in two ways: firstly, Janis's explanation is on the individual's level, it explains the behaviours of an individual in a group. Secondly the SIT's group justifications are about inter-group behaviours and attitudes, not intra-group ones. The groupthink theory cannot be proven within this study, but will be used as a possible explanation for inter-group conflicts in the data analysis.

Twitter as a platform

Twitter itself as a social media platform has great influence over its users and how the platform is and can be used. The tweets gathered in the data collection do not exist in a vacuum. The interactions between Twitter users are influenced by the platform itself. Certain aspects of Twitter as a platform should thus be considered in the analysis of the tweets: some tweets could be hyperbolic for comedic purposes or opinions could be voiced more bluntly than the users actually means to because of the limited characters.

Hyperbolism

The impact of the culture(s) of Twitter is worth mentioning, for they greatly set the tone for the tweets that are sent and twitter interactions. It is impossible to explain fully the culture of Twitter for it is not in the least homogenous and far too great a subject for this thesis to attempt to explain. However it should be stated that on Twitter it is not uncommon for users to make hyperbolic statements, often for the sake of humour or attention. The latter can be for many reasons, one of which being ‘trolling’: when an individual on the internet (better known as a ‘troll’) seeks to rile up other users, gain negative attention by posting inflammatory content on purpose. It is not completely possible to gauge the exact intent or tone of a tweet, therefore the possibility of hyperbolism creating an inflated sense of polarization should be considered.

Limited Characters

Twitter has a maximum of 280 characters per tweet. This limit of words per tweet greatly affects the level of nuance possible by the user. More than one tweet can be sent to convey the message fully (which is done often and goes by the name of ‘threads’). However the initial tweet is the one that is seen first and by a lot of users only; the initial tweet is the one that evokes (most) responses for other Twitter users. Besides, even through more tweets the character limit remains: tweets on Twitter are still limited in how they convey their messages.

Research Question

This study seeks to answer the following question:

How is online discourse on Twitter about canceling (someone) leading to group polarization?

Based on the literature I hypothesize the following results from the data analysis:

1. Based on the theories on polarization the following results are expected: more extreme opinions will be voiced by a user after getting positive feedback from other users or when similar opinions were voiced by other users (SCoT), arguments will be exchanged in the replies leading to more extreme ones (PAT) and, the individuals with very extreme opinions will not be persuaded at all by arguments that are contrary to the ones of their opinion (SET).
2. Based on the SIT I expect users who make category distinctions to engage in group justification behaviours by a) glorifying the in-group and/or b) derogating the out-group.

Methods

Design

A qualitative content analysis was chosen to find patterns of various reasonings in the exchanges between the Twitter users. The conversations between the users can only be properly analyzed if they are analyzed beyond the words written and include the symbolic meaning of the words as well. Content analysis includes the symbolic meaning of the texts, which allows for correlations and consequences of the interactions to be traced (Krippendorff, 1989).

The design of the analysis is created to understand the phenomenon of polarization but in a new context (on Twitter and with the subject being people being canceled). The theories on polarization form the basis for interpretation of the data, alongside the social theories on justifications and group behaviors. However they do not form a complete, integrated theory to be tested as would be the case with deductive designs. Neither does this study aim to purely use the data to form an independent new theory (induction). Therefore, the design is neither purely inductive or deductive, but actually abductive (Boeije & Bleijenbergh, 2019). The platform of Twitter was chosen because it is known to have lively debates on cancellations. No other social media platform is centered more around or enables conversations about topics such as the morality of celebrities and known-figures like Twitter. This makes the platform highly suitable for gathering data to answer the research question.

Study Population and Field Site

All Twitter users of the app/website Twitter.com who tweet in English or Dutch are included in the study population. However, Twitter has 187 million active users so not every user will be included but if a user discusses canceling someone they can be (Statista, 2021). The subject matter of the tweets is of highest importance in deciding whether or not to include a tweet. Targeting a specific subgroup of users to use tweets from would be impossible, because there is no clearly defined group of users who belong to the woke group. This is even more true for the supporter group because they contain multiple specific subgroups of supporters, dependent on the person they are supporting.

Data Collection

Data collection lasted 7 days and 717 tweets were collected. Tweet texts were copy-pasted into a data collection Word document with personal information removed to be used for further analysis. The content of the tweets could potentially include memes, videos, photos and gifs and were transferred (when possible) along with the text, because the media content can hold significant importance to understanding the messages. Gifs are not downloadable from Twitter and emoji's also could not be transferred to the Word document, but were described with words by the researcher. Tweets that replied to deleted tweets were not included because the meaning of the responding tweets were uninterpretable without context.

Only non-personal media were included, any forms of media that included personal information were not included but a memo was made of the specific tweet mentioning which form of media was used without divulging its contents. Personal media included any depictions of the user themselves or any markers of the person such as their face, body, home, workplace or other personal information. Reaction pictures of celebrities, pop culture characters, or others can therefore still be included. Posting such pictures is common practice on Twitter to convey the tone of the tweet by the user.

The relevant tweets were collected by searching for the selected keywords (listed below) twice daily, due to the different time zones in which the users are active and the high rate in which tweets accumulate. Tweets had to meet the following criteria to be included: The tweet that started the conversation had to include a) a reference to wokeness, woke culture, someone getting canceled, cancel culture or be an example of wokeness/cancel culture b) have at least 10 replies c) be written in English or Dutch. The starting tweet needed to have at least 10 replies in order to adequately be able to perform a discourse analysis. The reply tweets were then automatically included unless they contained sensitive/personal information or were deemed irrelevant (e.g. when a personal picture was included in the tweet or a tweet was made about a significantly different topic than cancel culture).

The following key terms have been chosen to find relevant tweets, because they are used often on Twitter by both the woke ingroup as the non-woke outgroup ('canceled', 'cancel').[3] Furthermore, terms being used often by the woke ingroup ('problematic' and 'toxic') were included to represent the views of the woke ingroup with regards to canceling. Terms being used often as a critique on woke culture were also included ('woke' and 'cancel culture') to represent the perspectives of the outgroup (from the woke perspective) with regards to canceling.

Ethics

This study was submitted to the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University and was approved (21-1760).

Privacy

Twitter is a public online platform, but its users do not post knowing or with the intention that their tweets could be used for research purposes. Anonymity has been ensured to protect the privacy of the Twitter users by copy-pasting the tweets without any user information.

Age

Twitter users must be thirteen years or older to have permission to create an account, making it possible for the users behind a portion of the analyzed data to be minors[4]. Twitter does not require proof of identification in the account set-up process, this in combination with the completely customizable profile makes it impossible to know whether the users are minors or not. As stated earlier the data included in the data analysis will be collected without any personal information. Any further actions do not seem possible to provide absolute

[3] These terms were found to be used often on Twitter by its users when a pilot of the data collection was done.

[4] https://help.twitter.com/en/managing-your-account/account-restoration

assurance that minors are not included in the study, due to the large number of users this study will be collecting tweets from.

Data Analysis

The data collected was analysed using NVivo qualitative content analysis. Relevant texts were coded starting with the ‘open coding’ phase, in which each selected text was cataloged as a node. Each node was coded as ‘TW1’, ‘TW2’, ‘TW3’ and so on. The next phase was the ‘axial coding’ phase in which the codes were integrated into larger categories (Boeije & Bleijenbergh, 2019). In this phase the goal was for a meaningful structure to start to form in which it would become clear which categories were related and how they differed from one another (Boeije & Bleijenbergh, 2019). For the axial coding phase a priori codes were created based on the literature. During the axial coding phase the accuracy of the existing categories (a priori and newly created) were intermittently checked to test if the existing codes were sufficient, needed to be changed or if new categories needed to be created. Any changes made during this process of redefining the categories were noted in memos in NVivo. Finally the selective coding phase started, in which underlying themes in the categories were analyzed to test this study’s hypothesis and answer the research question (Boeije & Bleijenbergh, 2019).

Results

Group polarization

Social Comparison Theory

According to the SCoT, positive social feedback and/or similar extreme opinions voiced by others can have an encouraging effect on an individual. This encouragement can give the individual the confidence to voice their privately held, more extreme opinion. Tweets 163-168 are a good example of positive feedback and sharing similar extreme opinions leading to more extreme opinions being voiced.

"TW163: What a shame! Unbelievable the censorship dat is going on everywhere.

TW164: It's worse than a shame. It's a disgusting unacceptable assault on everyone's personal freedom. That's why I am calling out to all my followers to go to Parler instead of Twitter.

TW165: I think I follow u on Parley. I will see straight away. And yes, a shame is an understatement. It is indeed unacceptable and we will have to keep fighting for our freedom, freedom of speech and autonomy.

TW166: Completely agreed. With this essential point there can be no compromises.

TW167: In the end everyone with an opinion will be thrown off of Twitter. The more you stand out, the quicker the opinion-police of Twitter will throw you out.

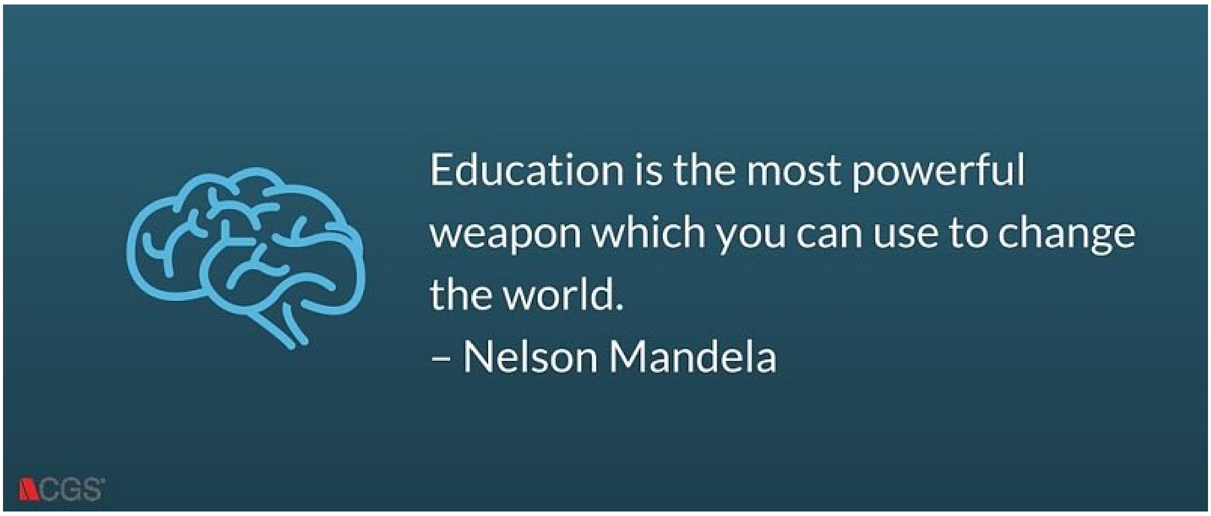
TW168: Indeed, and I refuse to subject myself to that."

At the beginning of the interaction the first user stated a mildly outraged statement, then the second user shared their similar but more extreme opinion. This leads the first user to share their more extreme opinion as well.

Stubborn Extremism Theory

Multiple conversations between users having extreme and opposing opinions showed no sign of flexibility. Once the extreme opinion was reached, no opinion was then seen to be amenable towards mitigation. The interaction between two users about education is an example of this:

TW21:"



Education is the most powerful weapon which you can use to change the world.
– Nelson Mandela

TW22: Indeed, but these days schools and uni's seem to think that leftish indoctrination is the same as education.
TW23: That's bs. Schools follow the curriculum. This shows you haven't been educated for quite some time. You should try it.
TW24: See the curriculum the UN is trying to make obligatory everywhere."

The writer of TW22 shares their extreme opinion on education, and when the writer of TW23 shares their opposing point of view the TW22 writer does not change their opinion even slightly (as seen in TW24).

"TW345: So @KiKa8118 has the luxury of judging the morality of its sponsors. I hope they get nothing from the VWS then. My monthly gift will keep coming of course. Well, until they no longer deem me worthy of donating of course.heeft de luxe om de moraliteit van donateurs te beoordelen.

TW346: No they are judging only the morality of the 'gift' itself, that's very different.
TW347: Do the sick children and their parents find that as well?"

Another example of an extreme opinion staying the same after another user voices their opposing opinion, confirming the stubborn extremism theory.

Persuasive Arguments Theory

As stated earlier the PAT could not be proven without conducting an experiment, which was not possible with this dataset. However the mitigating effects that the theory predicts when two opposing parties discuss their extreme opinions are demonstrated in the following discussion:

"TW667: There are people who have said and done bad things which were far less severe than criminal acts, but they have been near traumatized from the amount of online abuse they've received. A recent example: *link to a Youtube video about accountability*
TW668: Yet Lindsay is still employed, still makes money, still has fans and amazingly; wasn't "canceled"

Seems like a survivable situation
TW669: Silence, bot/troll/alt! *Picture of the Twitter user's replies to other user's tweets with the exact same message each time*[5]
TW670: In this example, Lindsay offended a subsection of her fans with a series of ignorant tweets. She also received a lot of hatred and attention from people outside her fan group for her actions. Cancelling isn't limited to "fans" in this sense, it's more like public shaming.
TW671: Lindsay was harmed by many individuals tweeting their disgust of her. "Cancel culture" can spawn remorseless dog-pilling of an individual, under the guise of just punishment for the crime. This is not true / ideal cancel culture (a community accountability system). This is abuse.
TW671: But overall I'm for individuals being held accountable for their speech and actions. Cancel culture is a productive step forward in the online community, if we cancel others with grace, respect and kindness (and if we can't do it that way, at least not directly abuse them)
TW672:Agreed. It is such a tough gray area. Sometimes people totally deserve time in the penalty box.

But, regardless everyone needs a refresher of the Golden Rule and stop with online threats and abuse."

The interaction clearly shows how two parties did (TW668 not included) grow more towards another with their opinions. This progression is in line with the predicted outcome of the PAT: when opposing parties discuss their opinions with each other their opinions will become more similar, more moderate.

Inter-group conflicts

Group justifications: in-group glorification

[5] The user who wrote this tweet accused the writer of TW667 of being a 'bot/troll/alt' meaning they thought the writer of TW667 was a robot, an automated Twitter account run by software, instead of a person

An interesting discovery was that cancel culture not only affects well-known figures, but within some Dutch communities on Twitter a member was canceled by being banned from Twitter. TW1: ‘Belangrijke mededeling: *naam twittergebruiker* is onlangs definitief van Twitter weg gecensureerd wegens gebruik van het woord "neger". Zij gaat waarschijnlijk verder op http://Parler.com. Ik zal mijn mening over dit wangedrag van Twitter voor me houden. U kunt het vast wel raden’, Group membership has in this situation become more tangible, because the defenders of the canceled and the canceled actually belong to the same group.

Although none of these tweets and others mentioned the in-group with actual distinctions, the lack of us distinctions being made outright were not seen as significant. Them distinctions were made and therefore and us distinction is implied. In-group glorifications were seen in multiple ways, but the two most prominent examples were the diminishment of in-group flaws and social change.

Diminishing in-group flaws

The user who got banned from Twitter was defended by members of their community: “Why is Sylvana Simons allowed to continue tweeting? And Gloria Wekker? Both racists. *name Twitter user* is not a racist and yet she was removed.
Stop your woke behavior, your identity politics and cultural marxism where the sun don’t shine.”(TW4). “I read tweets here on Twitter that go beyond absurde, yet the accounts get to stay. I have never seen *name Twitter user* do anything as bad.” (TW5).

In the users' defence of the canceled Twitter user they diminished the reasons for their removal from and by Twitter.

Social change

Multiple users used hashtags to try to get the canceled member of their community back on Twitter. For instance: “I miss @*name twitteruser* already. #bring*name twitteruser*Back.” (TW179). The user who wrote the first tweet (TW1) also wrote ‘Retweet as much as possible. Show that you are done with Twitter's censure."(TW2) to try to affect change through retweets.

Group justifications: out-group derogation

Multiple Twitter users made group distinctions in line with the SIT by Tajfel and Turner (1979). During many conversations users referred to woke people as being of another group 'They have completely lost their way. Did history not happen?"(TW273). Another user spoke of woke being a small group of people dictating what others are allowed to say. "Odd because that word is still here in The Netherlands in our dictionary! What happened in the world the last 5 years that has turned the world into an awful woke censorship world. A small flu dictates what we are allowed to say/write. And we still have no opposition to it!"' (TW28) The user who wrote this tweet later reiterated that they meant a small group of people, not a small flu, the user had made a typo.

Each time the out-group distinction was made the out-group was spoken of in a derogatory fashion. Examples of derogatory comments can, besides the generalized comments shown above, be categorized as personal attacks and discriminatory.

Personal attacks

During a debate on cancel culture, brought on by comments from comedian Katt Williams about cancel culture, some responses incited personal attacks multiple times. For example, in response to T493: 'I love Katt Williams, but this answer is horrifying. To me, cancel culture is not about a world where you can't say offensive, ignorant things. It's a world where you can't find redemption if you ever said something you regret. It's the inverse of let he who is without sin.' the user got replies such as 'Bullshit. NAME AN EXAMPLE of someone who got effectively "canceled" and couldn't find redemption. *emoji dark brown ear*' (T558), 'Cry more' (T560), and 'Lmao "horrifying." He said horrifying hahahahahahaha' (T571).

Discrimination

Multiple discriminatory comments were made by users such as: "There won't be any black Mozarts or Beethovens they don't have the skills. (TW250)" with as response: "Skills (means skills) which they do have, musical genius only in black music not in ours. (TW251). Another user wrote: “It is truly insane en something to cry about. Why is everyone crawling for those colored people of the world? Are we less than them? NO. Stand up for your past. It’s a part of our culture, our ancestors lived like that. Not us. So don’t change a thing.”(TW290).

The users who made the comments used language that put black people and people of color in distinctly different categories (‘they’ and ‘those people’). These comments could be seen as continuation of using ‘them’ distinctions, but since these comments were also specifically racist (which was not the norm within the data) they required their own sub-category.

Fake news, conspiracy theories, and calls to arms

Throughout readings of the data the following nodes were created as an addition to the a priori codes already made: 'Fake news', 'Conspiracy theories', and 'Call to arms'. They were added because multiple users made comments that did not fit the a priori codes that were made, but did seem to hold significant importance.

Firstly, at times the identity of a user was questioned, specifically if the user was a real person or not. Twice was a user suspected of being a bot: an automated Twitter account run by a software program. For instance:

TW57: Definitely.

But wasn’t this a fake account? I’m kind of lost.

TW58: Nee, it was just a serious account. I’ve been following her for a while now

TW59: Nope, it’s a dude.

TW60: Oh huh how do you know that then now I am lost too *questioning emoji*

Any suspicions about users being bots or distrust towards other information was coded under ‘fake news’.

Secondly, multiple users also proposed that the reasons behind certain events or processes were due to larger conspiracies. As one user stated: “It's a public secret that BigTeach, mostly its board, donated a lot of money to the Biden campagne.

The Project Veritas undercover video's were very clear as well: BigTeach is extremely leftist, not at all objective and censures anything they "deem necessary"." (TW150) The comments of the users themselves are not seen as true or untrue, simply that their statements implied that a larger conspiracy exists.

Lastly, multiple users spoke of action needing to be taken for social change and spoke to the other Twitter users in a call for action. One user wrote: "BARBARIANS ! DEFEND OUR EUROPEAN CIVILISATION !" (TW298). Another interaction showed a similar opinion shared:

> TW236: It is time that we started taking action. This madness has to end
> TW237: That won't be long now I think. It's all getting out of hand this way. It's not for nothing that they are warning about a civil war in France.
> TW238: A very good plan
> TW239: This is in no way different than the book burnings. In no way.
> TW240: Except perhaps that it's not about books.

The comments within this code could be seen as a continuation of group-justification: a desire to enhance the in-group's status through social actions. However these comments differed from the other tweets that were grouped as group-justification, because they spoke of a larger in-group. The in-group these users spoke of were of a national level, instead of a community level. Therefore, the comments differed greatly enough that they meritted their own code.

Discussion

In the conversations between users in which the two parties did not agree with each other the only group polarization theory that could be applied was the stubborn extremism theory. Most conversations between users holding opposing opinions, with at least one of them being extreme, lead to the extreme opinion not being mitigated. The conversation ended often with an equally strong or even stronger opinion on the matter discussed.

However, multiple conversations in which both opinions were strong, but somewhat moderated, and some that did start with at least one party stating an extreme opinion, were mitigated. Often these conversations petered out, without a clear ending of the debate. The conversation between users who shared the same or a similar opinion did in some conversations have an exacerbation effect on each other: ending with one or both of the users having a more extreme opinion voiced than before. It seems that the conversations on Twitter about cancel culture either stayed extreme, became more extreme or were somewhat mitigated. The reasons behind the differences in paths of polarization can be explained by the SET: the extreme pathway, the SCoT the more extreme pathway and the PAT the mitigated pathway.

Furthermore, the group polarisation on Twitter can be explained through group justifications and group-enhanced self-justifications. These forms of justifications could be underlying mechanisms that could explain the arguments had between the Twitter users. Once individuals start to identify with a group to a point where they see themselves as representatives of their group and/or where their self esteem is tied to their group, they could engage in group justification or group-enhanced self-justifications. The group member could justify its own actions or the actions of the group far more than if they had remained solely an individual. The data clearly showed that them distinctions were often made by Twitter users, therefore group memberships are in play. Both forms of justifications could have an inflammatory effect which could explain further why some of the conversations became more extreme through discussion. It is therefore possible that not only the social comparisons had an exacerbating effect on the opinions of the users, but the group memberships of the Twitter users as well.

Strengths and limitations

One of the most important discoveries made during the data collection process was the discovery of the Dutch communities who are very vocal about their dislike of cancel culture. Their input changed the trajectory of this thesis, because initially the woke individuals were hypothesised to be more vocal and would thus be the in-group. Yet the anti-cancel culture group was in multiple threads both a more tight knit community with a clear membership and more focussed on it’s dislike of cancel culture as a whole than that they were outraged over a specific person getting canceled. This new insight did not warrant new literary research or research model, but it did warrant a reappraisal of the assumptions that had been made. Within the Twitter debates the in-group was far more often the anti-cancel culture group (though not always). This group was also never solely angered by a person getting canceled, but more so by the capacity for someone to be canceled by others. Thus multiple assumptions made at the start of this thesis turned out to be wrong or askew.

This insight was made possible by collecting tweets without using a prompt, such as a tweet written by the researcher to gauge responses. This form of data collection allowed for natural conversations to be cataloged without interference from me as the researcher. The collected tweets were written by the users purely of their own accord, because those were the responses they wanted to reply with. Internal validity was therefore heightened, because the users/respondents were not as likely to respond in a socially desirable way towards the researcher, because they were not aware of research being done. Of course this covert manner of data collection brought a responsibility with it towards the users and their privacy. All tweets were carefully anonymized and only selected for collection when the criteria stated in 'Methods' were met. Important context of the certain tweets were written in memos(such as with TW28), as were any other significant notes of self reflection.

Another problem that the widely varying data of Twitter caused was that, because of the great range of topics discussed, it was hard to determine whether a statement was extreme or not. No clear parameters could be set for the statements, because the contexts of the tweets and different kinds of tweets kept changing per thread. Eventually all tweets were categorized as extreme when they were thought to be significantly different from the average or status quo opinion. This was determined by analyzing both the arguments used (e.g. violent comparisons, discriminatory or derogatory in nature) and the style in which they were written (e.g. hars language, angered emojis, more fiercely written that the other users)

Recommendation for future research

This study has highlighted multiple important insights that have implications on how polarising discours on Twitter can be and how it might work, but this study is only a small attempt at understanding a phenomenon about which virtually nothing is known. Further research is absolutely required on the topic of cancel culture, especially why cancel culture specifically appears to inflames people to such an extent that they call for justice. Theories on perceived injustices could possibly provide immensely interesting insights.

Implications and recommendations

One of the findings of the study was that the conversations on Twitter with like minded users with similar extreme opinions did lead to increased polarisation. This could be valuable information for Twitter and platforms like it: social media platforms could be contributing to polarisation. The other finding of the mitigating effects of discussions between opposing parties could also possibly be quite valuable. The different types of conversations being had on Twitter had vastly different outcomes with regards to how polarizing they were. If future research confirms these findings then Twitter and similar social media platforms would be advised to ask experts on how to manage the polarisation threats whilst still guarding its users rights of freedom of speech.

Conclusion

The conversations on Twitter about cancel culture have shown to be capable of polarizing or mitigating polarization on both the intra-group and inter-group level. The intra-group polarization and inter-group polarization appear to simultaneously affect the extent of group polarization. This study has demonstrated that multiple theories for both intra- and inter-group polarization could explain why cancel culture is a polarizing topic. However, to understand truly how the two processes work along each other, future research is required. The topic of cancel culture is still being fiercely debated, and by no means did the 717 tweets collected in this study capture the complete essence of why the topic has the ability to polarize Twitter users. Tweets collected from the final tweet (TW473) alone had thousands of replies, therefore it is likely that many more insights are to be found on this topic.

References

BBC News. (2020, June 9). Hartley Sawyer: The Flash actor fired over offensive tweets. Retrieved from https://www.bbc.com/news/entertainment-arts-52976556

Bishop, G. D., & Myers, D. G. (1974). Informational influence in group discussion. Organizational Behavior and Human Performance, 12 (1), 92–104. doi: 10.1016/0030-5073(74)90039-7

Boeije, H., & Bleijenbergh, I. L. (2019). *Analyseren in kwalitatief onderzoek* (3de ed.). Den Haag, Nederland: Boom Lemma.

BusinessInsider. (2020, June 3). HelloFresh severs partnership with Lea Michele after “Glee” co-star alleges the singer engaged in on-set bullying. Retrieved from https://www.businessinsider.nl/hellofresh-lea-michele-allegations-racist-bullying-2020-6?international=true&r=US

Change.org. (2020). Sign the Petition. Retrieved from https://www.change.org/p/dc-entertainment-remove-amber-heard-from-aquaman-2?redirect=false

Delbyck, C. (2020, November 11). “Fantastic Beasts” Set To Recast Johnny Depp With This Actor, And We’re Not “Mads” About It. Retrieved from https://www.huffpost.com/entry/fantastic-beasts-recasting-johnny-depp-mads-mikkelsen_n_5fac0dcfc5b68707d1fb4021

Editors of Merriam-Webster. (2017, November 7). What Does “Woke” Mean? Retrieved from https://www.merriam-webster.com/words-at-play/woke-meaning-origin

Editors of Merriam-Webster. (2021, January 21). “Getting Canceled” and “Cancel Culture”: What it Means. Retrieved from https://www.merriam-webster.com/words-at-play/cancel-culture-words-were-watching#:%7E:text=Uproxx%2C%2018%20Jan.-,2019,or%20promoting%20a%20writer’s%20works.

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. https://doi.org/10.1111/j.1751-9004.2007.00066.x

Janis, I. L. (1972). Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. Houghton Mifflin.

Krippendorff, K. (1989). Content analysis. In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth, & L. Gross (Eds.), International encyclopaedia of communication (pp. 403- 407). Oxford University Press.

Lerner, M. J., & Simmons, C. H. (1966). The observer’s reaction to the “innocent victim”: Compassion or rejection? Journal of Personality and Social Psychology, 4, 203-210

Mäs, M., & Flache, A. (2013). Differentiation without distancing. explaining bi-polarization of opinions without negative influence. PLoS ONE, 8 (11).

Myers, D. G., & Lamm, H. (1975). The Polarizing Effect of Group Discussion: The discovery that discussion tends to enhance the average prediscussion tendency has stimulated new insights about the nature of group influence. American Scientist, 63(3), 297–303.

Myers, D. G. (1982). Polarizing Effects of Social Interaction. In H. Brandstätter, J. H. Davis, & G. StockerKreichgauer (Eds.), Group decision making. London: Academic Press

NOS. (2020, July 8). J.K. Rowling vraagt met 149 anderen om einde “cancel-cultuur”: “Debat onmogelijk.” Retrieved from https://nos.nl/artikel/2340001-j-k-rowling-vraagt-met-149-anderen-om-einde-cancel-cultuur-debat-onmogelijk.html

Oakes, P. J. (1987). The salience of social categories. In J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, S., & Wetherell, M. S. Rediscovering the Social Group: A Self-categorization Theory (pp. 117–141). Oxford, UK: Blackwell.

Oakes, P. J., Turner, J. C., & Haslam, S. A. (1991). Perceiving people as group members: The role of fit in the salience of social categorizations. British Journal of Social Psychology, 30, 125–144

Oakes, P. J., Haslam, S. A., & Turner, J. C. (1994). Stereotyping and Social Reality. Oxford, UK: Blackwell.

Sarkisian, J., & Ntim, Z. (2021, March 25). A complete timeline of Johnny Depp and Amber Heard’s tumultuous relationship. Retrieved from https://www.insider.com/johnny-depp-amber-heard-relationship-timeline-2020-7

Spears, R., Oakes, P. J., Ellemers, N., & Haslam, S. A. (Eds.) (1997). The Social Psychology of Stereotyping and Group Life. Oxford, UK: Blackwell.

Statista. (2021, February 9). Countries with the most Twitter users 2021. Retrieved May 15, 2021, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/#:%7E:text=Global%20Twitter%20usage,former%20U.S.%20president%20Barack%20Obama.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), The Social Psychology of Intergroup Relations (pp. 33–47). Monterey, CA: Brooks/Cole.

Tajfel, H. (1981). Social stereotypes and social groups. In J. C. Turner & H. Giles (Eds.), Intergroup Behaviour (pp. 144–167). Oxford, UK: Blackwell

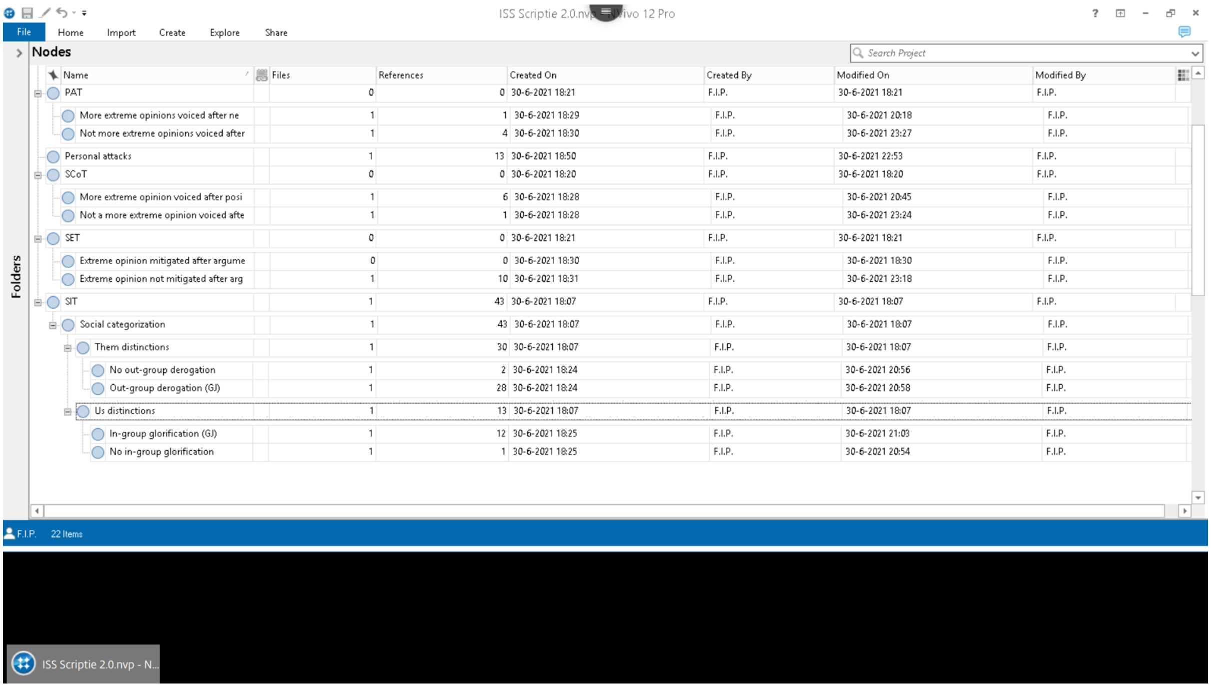
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). Rediscovering the social group: A self-categorization theory. Oxford: Blackwell.

Turner, M. A., & Smaldino, P. E. (2018). Paths to Polarization: How Extreme Views, Miscommunication, and Random Chance Drive Opinion Dynamics. *Complexity*, *2018*, 1–17. https://doi.org/10.1155/2018/2740959

Turner, M. A., & Smaldino, P. E. (2020). *Stubborn extremism as a potential pathway to group polarization*. Retrieved from https://www.researchgate.net/publication/343365284_Group_polarization_via_stubborn_extremism

Vinokur, A., & Burstein, E. (1974). Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach. Journal of Personality and Social Psychology, 29 (3), 305–315. doi: 10.1037/h0036010

Appendix: Code tree



Nodes

Name	Files	References	Created On	Created By	Modified On	Modified By
PAT	0	0	30-6-2021 18:21	F.I.P.	30-6-2021 18:21	F.I.P.
More extreme opinions voiced after ne	1	1	30-6-2021 18:29	F.I.P.	30-6-2021 20:18	F.I.P.
Not more extreme opinions voiced after	1	4	30-6-2021 18:30	F.I.P.	30-6-2021 23:27	F.I.P.
Personal attacks	1	13	30-6-2021 18:50	F.I.P.	30-6-2021 22:53	F.I.P.
SCoT	0	0	30-6-2021 18:20	F.I.P.	30-6-2021 18:20	F.I.P.
More extreme opinion voiced after posi	1	6	30-6-2021 18:28	F.I.P.	30-6-2021 20:45	F.I.P.
Not a more extreme opinion voiced afte	1	1	30-6-2021 18:28	F.I.P.	30-6-2021 23:24	F.I.P.
SET	0	0	30-6-2021 18:21	F.I.P.	30-6-2021 18:21	F.I.P.
Extreme opinion mitigated after argume	0	0	30-6-2021 18:30	F.I.P.	30-6-2021 18:30	F.I.P.
Extreme opinion not mitigated after arg	1	10	30-6-2021 18:31	F.I.P.	30-6-2021 23:18	F.I.P.
SIT	1	43	30-6-2021 18:07	F.I.P.	30-6-2021 18:07	F.I.P.
Social categorization	1	43	30-6-2021 18:07	F.I.P.	30-6-2021 18:07	F.I.P.
Them distinctions	1	30	30-6-2021 18:07	F.I.P.	30-6-2021 18:07	F.I.P.
No out-group derogation	1	2	30-6-2021 18:24	F.I.P.	30-6-2021 20:56	F.I.P.
Out-group derogation (GJ)	1	28	30-6-2021 18:24	F.I.P.	30-6-2021 20:58	F.I.P.
Us distinctions	1	13	30-6-2021 18:07	F.I.P.	30-6-2021 18:07	F.I.P.
In-group glorification (GJ)	1	12	30-6-2021 18:25	F.I.P.	30-6-2021 21:03	F.I.P.
No in-group glorification	1	1	30-6-2021 18:25	F.I.P.	30-6-2021 20:54	F.I.P.

Appendix: Interdisciplinariteit Reflectie:

1. In hoeverre draagt het gebruik van theoretische inzichten uit meerdere wetenschappelijke disciplines bij aan het begrip van het te onderzoeken probleem?

Interdisciplinariteit is van grote waarde in dit onderzoek, omdat het probleem: het polariserende effect van woke discourse een complex probleem is. Één discipline zou niet toereikend zijn. Door de inzichten van meerdere disciplines bijeen te brengen kan het probleem van meerdere invalshoeken bekeken worden en kunnen meer aspecten van het probleem geanalyseerd worden. Des te meer aspecten van het probleem doorgrond worden des te beter wordt het probleem in het geheel doorgrond en wordt een oplossing meer mogelijk.

2. Als het gebruik van theoretische inzichten uit meerdere disciplines zinvol is: *welke* disciplines zijn dit, en waarom is het zinvol om gebruik te maken van theoretische inzichten uit juist *deze* disciplines?

Sociologie en psychologie, omdat de inzichten in gedrag van groot belang zijn in het begrijpen hoe polarisering plaats kan vinden. Om te begrijpen waarom het onderwerp van canceling polarisering kan veroorzaken zijn mentale processen begrijpen van groot belang.

3. In hoeverre kan het gebruik van meerdere wetenschappelijke onderzoeksmethoden bij het onderzoeken van het probleem leiden tot meer inzicht in het probleem? En als het gebruik van meerdere wetenschappelijke onderzoeksmethoden naar verwachting zal leiden tot meer inzicht in het probleem: *welke* onderzoeksmethoden zijn dit, en waarom is het gebruik van juist *deze* onderzoeksmethoden zinvol?

   Meerdere wetenschappelijke onderzoeksmethoden gebruiken als een vorm van triangulatie leidt tot hogere validiteit wat het onderzoek sterker maakt. Daarnaast kunnen sommige aspecten van een probleem soms niet met één meetinstrument adequaat gemeten worden en is een uitbreiding van onderzoeksinstrumenten/methoden nodig. In dit onderzoek zou een focusgroep tussen woke mensen en supporters van een bepaalde gecancelde persoon een uitermate interessante uitbreiding zijn. De polarisatie mechanismen zouden dan getest kunnen worden, maar dit zou wellicht wel online uitgevoerd moeten worden omdat anders zou het onderzoeksdesign te ver van het Twitterdesign af gaan.

4. In hoeverre kan het gebruik van meerdere analytische niveaus bij het onderzoeken van het probleem leiden tot meer inzicht in het probleem? En als het gebruik van meerdere analyseniveaus naar verwachting zal leiden tot meer inzicht in het probleem: *welke* niveaus zijn dit, en waarom is de gezamenlijke analyse van juist *deze* niveaus zinvol?

In dit onderzoek lijkt meerdere analytische niveaus niet te leiden tot meer inzichten voor het probleem.