FROM INDIVIDUALS TO PARTISANS: THE ROLE OF SOCIAL CATEGORIZATION IN POLITICAL POLARIZATION

Carsten W. Sander

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FROM INDIVIDUALS TO PARTISANS: THE ROLE OF SOCIAL CATEGORIZATION IN POLITICAL POLARIZATION

Dissertation zur Erlangung des Doktorgrades an der

Universität Hamburg,
Fakultät für Psychologie und Bewegungswissenschaft,
Institut für Psychologie

vorgelegt von

Carsten W. Sander

Hamburg, 2026

Tag der mündlichen Prüfung

03. September 2026

Promotionsprüfungsausschuss

Vorsitz: Prof. Dr. Tania Lincoln

1. Dissertationsgutachten: Prof. Dr. Juliane Degner

2. Dissertationsgutachten: Prof. Dr. Lena Frischlich

1. Disputationsgutachten: Prof. Dr. Nale Lehmann-Willenbrock

2. Disputationsgutachten: Prof. Dr. Jenny Wagner

Acknowledgements

I want to sincerely thank Prof. Dr. Juliane Degner for making this not just an extremely insightful and enriching but also genuinely enjoyable experience, for teaching me to think like a researcher, for granting me more intellectual autonomy than I probably deserved, for bearing with me whenever my academic stubbornness was giving her migraines, for having a great moral compass and a terrific sense of humor (irony detection aside), and for tolerating office memes that were not rarely at her expense.

I also thank my deskmate-turned-friend-for-life Jana Mangels, who I should probably have listed as co-supervisor somewhere and who never felt—or, more realistically, never showed—any annoyance when I disrupted her focus with anything from a spontaneous half-baked research idea to a full-blown emotional crisis. While we arguably distracted each other a lot, the distractions were as uplifting and, at times, productive as they come.

But the positive distractions did of course extend past the boundaries of our shared office, leaving me with deep gratitude for my colleagues Joelle Flöther, Anna Huang, and Feline Becker, as well as Stella Fingas, who technically belongs to a different department but clearly never got the memo. Through hilariously unsuccessful sailing maneuvers, missed pun potential of legendary proportions, otherworldly active listening skills, or office visits so frequent and welcome that I never once felt the need to return them, they made sure that our lab is the nest I will never want to leave.

For steadying me during the ups and downs of academic life, my deepest gratitude also goes to friends, loved ones, and family members, whether they were there for all of it or only for some. This dissertation—including any typos, formatting errors, flawed reasoning, accidental plagiarism, or harmful societal implications that may or may not be discovered after publication—is your success as much as mine. But jokes aside, I am forever grateful for your emotional and, where applicable, financial support. Lastly, speaking of financial support, I gratefully acknowledge the doctoral scholarships granted by the Studienstiftung des deutschen Volkes and the University of Hamburg.

Abstract

Political polarization is increasingly viewed not as a process of diverging political opinions but of rising levels of partisan prejudice. This so-called affective polarization is argued to influence how members of different political groups relate with one another, with potentially severe consequences for social cohesion and democratic functioning. While many researchers have therefore set out to find ways of reducing partisan prejudice, they have so far neglected a crucial prerequisite for most of the presumed negative consequences of partisan prejudice: social categorization. Prejudice toward the political outgroup can only affect how we relate to outgroup members when they are identified as such. Given that political categories are not tied to any invariant perceptual features, the present dissertation asks how individuals in their day-to-day lives come to perceive each other as partisans and to what extent this tendency is malleable. The first dissertation article showed across a series of five experiments that individuals inferred a wide range of political categories from behavioral information (e.g., expressions of political opinions). Moreover, these inferences were shown to occur spontaneously and to be robust in that they persisted even when alternative situational reasons for the behaviors were provided. The second article is premised on the idea that these robust categorizations are grounded in the behavioral cues being perceived as highly diagnostic (i.e., uniquely tied to the inferred category), and showed across three experiments that perceptions of diagnosticity can be reduced through counterstereotypical exemplars highlighting the occurrence of the cue outside of the inferred category. The third article builds on findings suggesting that partisan prejudice strongly depends on entitativity, that is, the degree to which category members are perceived as a coherent unit. It showed that the use of group labels—a known driver of entitativity—in social media posts predicted expressions of partisan prejudice in the respective comments. The dissertation concludes with a discussion of how the constructs of the three articles relate to each other and why they should be considered a driver and potential leverage point of polarization, rather than a mere epiphenomenon of it.

Table of Contents

CHAPTER 1

General Introduction 1

CHAPTER 2

First Article · Investigating Mechanisms of Political Polarization: Perceivers Spontaneously Infer Ideological Categories from Other People's Behavior 27

CHAPTER 3

Second Article · From Partisans to Individuals: Lowering an Opinion's Diagnosticity via Counterstereotypes Can Reduce Categorization Based on That Opinion 43

CHAPTER 4

Third Article · Constructing the Enemy: Noun and Generic Party Labels Predict Partisan Prejudice in Social Media Comments 71

CHAPTER 5

General Discussion 105

SUPPLEMENTAL MATERIALS I

First Article 143

SUPPLEMENTAL MATERIALS II

Second Article 177

SUPPLEMENTAL MATERIALS III

Third Article 201

APPENDIX

Declarations 227

CHAPTER 1
General Introduction

CHAPTER 1

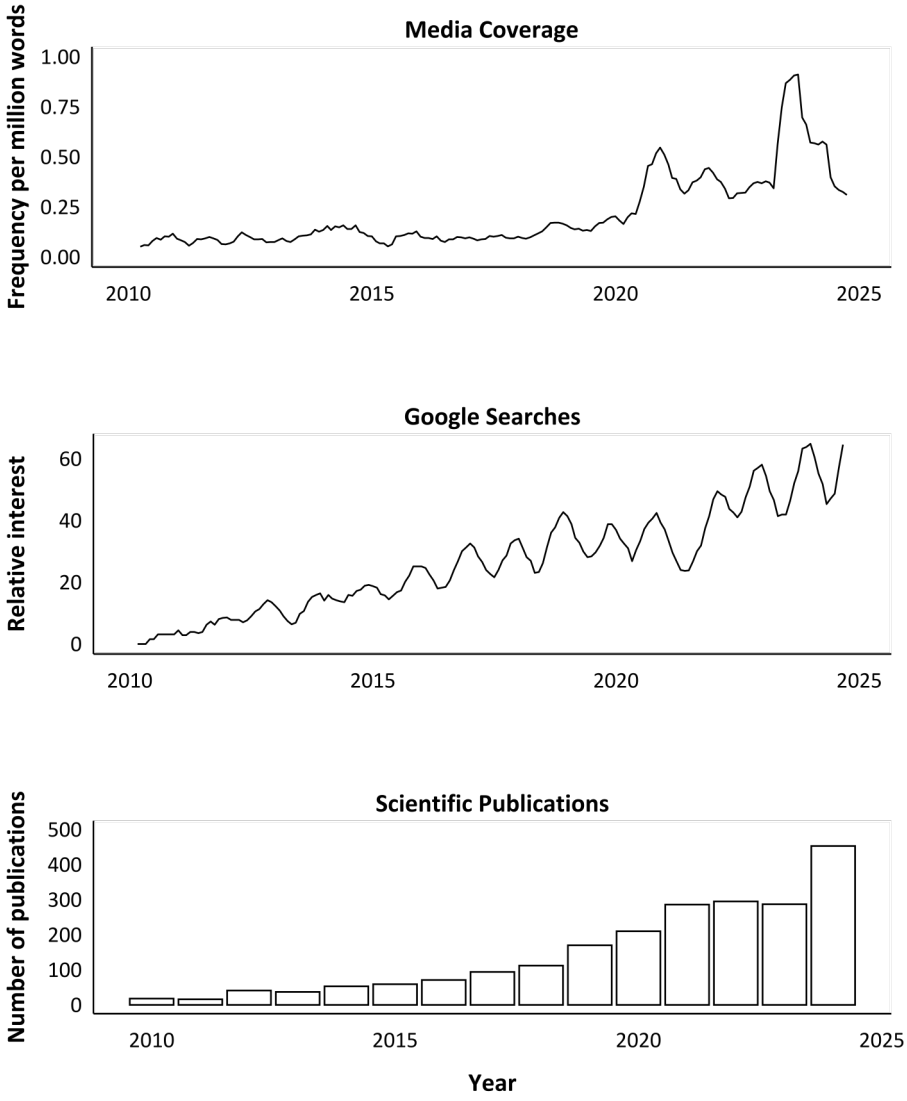
Sensing Polarization

Political pundits and the broader public alike seem to sense that society is becoming increasingly divided. In the United States—the poster child of political polarization—public attention to the phenomenon has increased measurably in recent years: Intensifying media coverage of the topic as well as more frequent Google searches suggest that public interest on political polarization is increasing, and a growing number of scientific publications per year suggests that scholarly attention is also on the rise (see Figure 1). Other countries appear to mirror these trends: In Germany, where some studies of this dissertation were conducted, 77 percent of the public believes that opinions in society are increasingly diverging, and 70 percent feels that public debates are becoming more emotional and less factual (Wissenschaft im Dialog, 2025).

Despite this widespread sense that polarization is intensifying, demonstrating it empirically has proven rather challenging. Early conceptualizations focused on the distribution of individuals' policy positions, yet these approaches often failed to reveal the expected patterns among the mass public. In response, researchers shifted their attention to the social dimensions of polarization, that is, how negatively supporters of different political parties or ideological camps view one another. This reconceptualization of polarization as increasingly negative attitudes toward political outgroups aligns more closely with public intuitions. While many scholars argue that this type of polarization is very consequential, not least because it negatively affects how citizens relate to those that they perceive as political outgroup members (e.g., Finkel et al., 2020; Iyengar et al., 2019), previous research has left a critical question unaddressed: How do people come to perceive one another as members of political outgroups in the first place? In my dissertation, I investigate the socio-cognitive mechanisms underlying the tendency to see fellow citizens as partisans rather than individuals and discuss to what extent polarization may be driven by these mechanisms.

Figure 1

Indices of Rising Interest in Polarization



Note. The upper panel shows the frequency of the term "political polarization" in the Corpus of News on the Web (NOW; Davies, 2016), which contains 23.4 billion words of English-language text from web-based newspapers and magazines. Only U.S. sources and instances in which the term functions as a noun are included. A centered six-month moving average was computed and plotted over time. The middle panel displays a centered six-month moving average of relative search interest in the term "political polarization" based on Google Trends data for the United States. To account for sampling variability in Google Trends data (Gummer & Oehrlein, 2025), ten separate samples were collected, and the median value at each time point is shown. The lower panel presents the results of a bibliometric analysis based on Web of Science data, showing the number of publications with the topic "political polarization" over time.

CHAPTER 1

Ideological Polarization

Previously, polarization was predominantly understood as an issue-based or ideological phenomenon: a trend in the distribution of people's policy positions whereby the electorate divides into two opposing camps with increasingly non-overlapping opinions. For the United States, which serves as the primary reference case throughout this dissertation given the density of available empirical literature, considerable scientific debate has centered on whether this so-called ideological polarization has actually increased. Gauging its extent is difficult because measurement fundamentally depends on which opinions are included in the analysis. Some researchers restrict their analyses to survey items featured at all investigated timepoints to control for differences between items (e.g., Ojer et al., 2025), whereas others include less frequently used items to account for changes in issue relevance over the years (e.g., Baldassarri & Gelman, 2008; Kozlowski & Murphy, 2021). Additionally, the debate is grounded in conceptual ambiguities: A decreasing overlap in opinions can theoretically result from ideological divergence, whereby opinions become more extreme in opposing directions (Fiorina & Abrams, 2008), or from alignment, whereby opinions (and identities) become more intercorrelated without necessarily becoming more extreme (Abramowitz & Saunders, 2008).

With regard to ideological divergence, there is little evidence of substantial increases among the American mass public. Distributions of political opinions appear to remain largely unimodal (Fiorina & Abrams, 2008; Lelkes, 2016). For partisans (i.e., individuals identifying as Republicans or Democrats), on the other hand, there is evidence of increased ideological divergence over the last few decades (Enders, 2021; Lelkes, 2016; Ojer et al., 2025), though considerable overlap between partisan groups remains (Enders, 2021; Lelkes, 2016). With regard to alignment, initial evidence suggested that increases were confined to the relationship between opinions and identities rather than the interrelations of different opinions. That is, opinions appeared to increasingly map onto ideological identities (e.g., conservative vs. liberal) and party identities (e.g., Republican vs. Democrat), whereas issue alignment—the degree to which different opinions correlate with one

another and form ideologically constrained belief systems—seemed to increase merely among strong partisans (Baldassarri & Gelman, 2008). The mass public's opinions, on the other hand, appeared to retain relatively low levels of intercorrelation (Baldassarri & Gelman, 2008; Lelkes, 2016). However, more recent data suggest that the rate at which alignment is increasing has accelerated, even for the mass public. For instance, Kozlowski and Murphy (2021) found a strong increase in issue alignment from 2004 to 2016, both for opinions within and between different issue domains (e.g., economic, civil rights). Analyses by Hare (2022) using a different methodology confirm this trend across multiple issue domains including abortion, environmental regulations, health insurance, and immigration (but see Park, 2018). Though issue alignment is consistently highest among politically sophisticated individuals, the increases also appear among the politically less sophisticated (Hare, 2022; Kozlowski & Murphy, 2021).

Research on European democracies shows broadly comparable patterns, except for the recent increase in alignment observed in the United States. For instance, Gestefeld et al. (2022) found no Europe-wide trend in ideological divergence across 33 European countries. For Germany, Hebenstreit (2023) found no evidence of ideological divergence except for the issue of immigration. With regard to party alignment, Cely (2025) reported a left-right asymmetry for European countries, with supporters of left-wing parties showing higher levels of ideological constraint compared to supporters of right-wing parties. Although a comparison between 2008 and 2016 revealed no general trend toward higher party alignment, people identifying with culturally conservative parties showed a strong increase, suggesting that this asymmetry has vanished for party alignment on sociocultural issues. A study by Caldwell et al. (2025) focusing only on climate policy and economic issue positions found significant increases in party alignment over recent decades across Anglophone and Western European countries, but not elsewhere. The authors also found increases in issue alignment for these two issue domains. In contrast, Warncke (2025) examined eight different issues and found only large between-country differences but no meaningful changes over time (2002-2020) in average issue alignment across 38 European countries (though the inclusion of

Eastern European countries may have obscured increases in Western European ones). Despite the slightly broader range of issues used by Warncke (2025), an even broader range would be favorable for estimating trends in overall issue alignment.

Taken together, apart from more recent trends in the US, different aspects of ideological polarization are either relatively stable or have increased only mildly or for specific issues across many democracies. Generally, ideological polarization appears strongest and shows higher increases among partisans or politically sophisticated individuals (e.g., those with high political knowledge or education), and when considering party or ideological alignment rather than issue alignment. Although observed changes in ideological polarization were predominantly increases rather than decreases, the magnitude of these increases seems too small—and in the case of the US too recent—to fully explain the pervasive sense that polarization is on the rise.

Affective polarization

Given the seemingly insufficient ideological basis of the apparent polarization, researchers began to investigate polarization as a social intergroup phenomenon. In this view, citizens identify with a party, making it their ingroup, and perceive the opposing party as an outgroup. Social Identity Theory (Tajfel & Turner, 1979) suggests that the mere identification with a group should lead to a more favorable evaluation of the ingroup compared to the outgroup, especially in intergroup contexts that are perceived as competitive. Scores of research have confirmed this idea, showing that even the identification with virtually meaningless groups can lead to a host of systematic biases (for a review see Dunham, 2018). As the relationship between political parties is by definition competitive, identification as a Republican or Democrat could therefore lead to considerable polarization in group evaluations and related outcomes even in the absence of severe ideological disagreement (Iyengar & Westwood, 2015). The divergence of inparty and outparty evaluations is called affective polarization and has, in contrast to ideological polarization, indeed dramatically increased in the US since 1990, driven predominantly by more negative affect toward the outparty (Iyengar et al., 2012). Although most available longitudinal data is based on evaluations of parties

rather than party supporters, it is reasonable to assume that similar trends have occurred for the latter, given that both evaluations are highly correlated and conceptually related measures like social distance or trait rating scales aimed at partisans mirror the trend in party evaluations (Iyengar et al., 2012).

Comparative studies of affective polarization across democracies reveal no uniform global trend and instead point to divergent country-level developments (Boxell et al., 2024; Garzia et al., 2023). Evidence for Germany is somewhat contradictory: Boxell et al. (2024) reported a pronounced decline in affective polarization over recent decades, whereas Garzia et al. (2023) identified a clear increase. These contrasting findings appear to stem largely from methodological differences. For instance, Boxell et al. (2024) excluded respondents from East Germany, while Garzia et al. (2023) included them. In addition, the two studies employed different operationalizations of affective polarization. In multiparty systems such as Germany, citizens may often lack a single, clearly defined ingroup and instead hold multiple party attachments of varying strength (Rothers, 2025). Reflecting this complexity, Wagner (2021) proposed measuring affective polarization as the dispersion of like-dislike ratings across parties (potentially weighted by parties' vote shares) without assuming exclusive party identification. Garzia et al. (2023) adopted this approach and analyzed the full electorate, whereas Boxell et al. (2024) assumed a single in-party and restricted their analysis to explicit partisans. But perhaps most importantly, Garzia et al. (2023) examined a substantially longer time span, incorporating the years from 1961 to 1977, which were characterized by exceptionally low levels of affective polarization. Because these early observations are missing from Boxell et al.'s (2024) data, the two studies arrived at opposing trend estimates. Given these inconsistencies, it may be preferable to focus on shorter time frames. Analyses using more fine-grained temporal data suggest that affective polarization in Germany declined between 1990 and 2008 but has risen steadily since, reaching high—though not entirely unprecedented—levels by 2019 (Borbáth et al., 2023; Harteveld & Wagner, 2023). Thus, when we do not reduce polarization to diverging ideology, it appears to be indeed increasing in both the US and more recently also in Germany.

While affective polarization is typically operationalized via partisan prejudice—specifically, the gap in in- versus out-party affect—it may reflect a richer underlying phenomenon. One aspect of this is partisan stereotyping: Americans hold measurably distorted beliefs about both partisan groups, overestimating their opinion extremity and homogeneity (Ahler, 2014; Dias et al., 2025; Druckman et al., 2022; Levendusky & Malhotra, 2016; Westfall et al., 2015; Yudkin et al., 2019), the overlap between partisanship and other social characteristics such as income or sexual orientation (Ahler & Sood, 2018), and the extent to which members of their out-party view members of their in-party negatively (negative meta-perceptions; Lees & Cikara, 2020; Moore-Berg et al., 2020; Ruggeri et al., 2021). Consistent with research establishing the causal interdependence of stereotypes and prejudice (Phills et al., 2020), these misperceptions have been shown to have a causal effect on partisan prejudice (Ahler & Sood, 2018; Druckman et al., 2022; Myers & Hvidsten, 2024). Another aspect of affective polarization is how individuals behave toward one another. People high in partisan prejudice show reduced empathy, trust, and willingness to engage with out-partisans relative to co-partisans (Broockman et al., 2023; Carlin & Love, 2013; Simas et al., 2020), and as a result out-partisans are actively avoided across a range of consequential domains such as online dating (Huber & Malhotra, 2017), roommate selection (Shafranek, 2021), hiring (Gift & Gift, 2015; Inbar & Lammers, 2012), and even intrametropolitan migration (Gimpel et al., 2025). Measures of affective polarization may therefore be better understood not as mere indices of negative out-party affect, but as proxies for broader societal conflict, reflecting exaggerated intergroup stereotypes, prejudice, and real-world discrimination.

Levels of affective polarization have been linked to several negative societally relevant outcomes such as reduced institutional and social trust (Torcal & Thomson, 2023), reduced democratic support (Janssen & Turkenburg, 2025), increased political violence (Piazza, 2023), democratic backsliding (Orhan, 2022), and even decreased self-rated health (Nelson, 2022). For this reason, a lot of scholarly attention is directed at the goal of reducing affective polarization (e.g., Hartman et al., 2022; Voelkel et al., 2024). However, reducing affective polarization may not be the

only way to reduce its negative consequences. In the following I argue that this may also be achieved by reducing the number of situations in which people are aware of each other's political affiliations.

Social Categorization

While research on political polarization has benefited greatly from adopting a social identity perspective, it can also profit from the literature on social cognition. In particular, much existing work overlooks a crucial prerequisite for many of the presumed consequences of affective polarization: social categorization, the cognitive process through which people sort themselves and others into social groups. The adverse effects associated with affective polarization—such as stereotyping, contact avoidance, distrust, discrimination, or reduced empathy—can only occur once targets have been identified as outpartisans. This distinction is well established in the social cognition literature, which differentiates between category activation and the application of stereotypes or prejudice (e.g., Macrae & Bodenhausen, 2000; Roth et al., 2019). Categorical representations of political outgroups—including partisan stereotypes and evaluative biases—can only be applied to a target if it has first been categorized. Yet most polarization research concentrates on the content of partisan representations while paying little to no attention to how individuals are identified as in- or outpartisans in everyday contexts.

Although this distinction may initially appear pedantic, it is theoretically consequential. In many extensively studied intergroup domains—such as age, gender, or racialized group membership—category activation is often taken for granted because the cues used for categorization are highly salient and continuously available. In everyday interactions, people routinely extract such categorical information from facial features alone (Colombatto et al., 2021; Mandalaywala & Legaspi, 2023). Names may also convey (probabilistic) information about gender, age, racialized group membership, or religion. These cues are relatively stable, difficult to conceal, and largely outside individuals' immediate control.

Partisanship and political ideology differ fundamentally in this respect. Faces or names typically do not reveal whether someone is politically left- or right-leaning or supports a particular

party. Instead, political categories must be inferred from cues that are usually under individuals' conscious control. Such cues may include the open display of political identity symbols (e.g., yard signs, bumper stickers, MAGA hats, or pride flags), deliberate expression of specific political opinions (e.g., advocating abortion bans or demanding stricter gun control measures), but also more subtle cues such as the adherence to polarized social norms (e.g., including preferred pronouns in an email signature) or lifestyle signals such as food choices, brands, clothing, and hobbies. These cues become very potent when they can very frequently be observed by other people. The COVID-19 pandemic provides an illustrative example. Wearing versus not wearing masks in public spaces was an extremely visible behavior and as research documenting a substantial partisan gap in adherence to mask mandates in the United States shows, it became a relatively reliable indicator of political ideology (Boykin et al., 2021; Powdthavee et al., 2021; see also Bunosso et al., 2022). By serving as a partisan cue, mask-wearing thus allowed the actualization of numerous categorization effects (e.g., heated arguments, avoidance behaviors) that would not have manifested otherwise. Another example are linguistic practices, which are also very potent due to their high observability and frequent use. In Germany, for example, the use of so-called gender-fair language—linguistic practices intended to explicitly include all genders, such as avoiding generic masculine forms or using special spellings to mark gender diversity—has become a highly affectively polarized issue (Wissenschaft im Dialog, 2025). Thus, in contrast to many other categories, the extent to which political categories can be inferred depends on whether individuals choose to display visible partisan or ideological cues.

For this reason, the inferability of political categories is much more malleable than that of many other categories and can change over time. Whether or not the overall prevalence of categorization cues is on the rise or in decline is hard if not impossible to gauge. But at least on social media, where this may be easiest to study, many potential categorization cues were shown to be increasingly prevalent in recent years: For instance, in U.S. Americans' Twitter/X profiles the prevalence of expressions of preferred pronouns has risen considerably (Tucker & Jones, 2026), as

has the prevalence of LGBTQ signifiers such as pride flag emojis (Jones & Cisternino, 2024) and other political keywords that either explicitly (e.g., “Progressive”, “Republican”, or “Communist”) or implicitly (e.g., “Blue lives matter”, “Feminist”, “MAGA”) signal political identity (Rogers & Jones, 2021). A general increase in the prevalence of categorization cues thus appears at least plausible. Because categorization enables the application of negative outgroup stereotypes and evaluations and is much more malleable in the political compared to other domains, it is an important factor in polarization dynamics that warrants systematic theoretical and empirical attention.

Overview

In my dissertation I investigated the role of social categorization in political polarization. In the following I provide an overview of the empirical Chapters 2, 3, and 4, each of which corresponds to one of the three dissertation publications. Chapter 5 then concludes by discussing potential synergies between the articles’ core concepts and places them in the larger societal context.

First Article

I have argued that people usually cannot infer partisanship from someone’s name or facial features and therefore need to use behavioral cues such as expressions of opinions, the use of identity symbols, or the adherence to polarized social norms. While it may seem self-evident that such an inferential process would be necessary to arrive at political categorizations, it is not evident that this process actually occurs in real-life. Subjectively, we may all be familiar with the feeling that a person suddenly becomes likable or unlikable to us because of an expressed opinion or another political cue. But does this automatically mean that we have categorized them? Some political scientists would argue it does not. Rather than assuming that cues, specifically opinions, elicit categorization, some researchers assume that the opinions themselves, or our disagreement with them, are at the core of partisan animosity (Orr & Huber, 2020; Webster & Abramowitz, 2017). As partisanship and opinions mutually inform each other, it can be difficult to experimentally tease apart the two processes’ impact on partisan animosity (but see Dias & Lelkes, 2022). Chapter 2 of the dissertation adds to this discussion, by investigating the process of categorization more directly,

rather than via its downstream consequences. Through several preliminary studies, we identified commonplace observable behaviors (e.g., using anti-nuclear power stickers, criticizing the gender pay gap, complaining about gender-fair language) that German participants—when explicitly prompted to do so—associated with particular political groups (e.g., green, feminist, conservative). Using indirect impression formation paradigms from the impression formation literature, we then showed that these categories are also inferred spontaneously, that is, when participants were not explicitly prompted to categorize or form person impressions. Moreover, we showed that these categorizations were quite robust, to the extent that they persisted when alternative situational reasons for the cues were provided.

Second Article

For the second article we focused on the question why certain cues elicit such robust categorizations and whether the degree to which these cues afford categorizations can be reduced. Given our focus on behavioral cues, we looked for answers in the literature on trait inferences, which similarly investigates behavior-based categorization processes. This literature ties categorical inferences and the robustness thereof to the concept of diagnosticity—the degree to which a cue is likely to be observed in one category and unlikely in another (Gawronski, 2004; Skowronski, 2002). In other words, cues are assumed to elicit robust categorizations to the extent that they are exhibited by members of one category but not by members of another. To illustrate, a phenomenon that can be explained this way is the tendency to give more weight to negative behaviors when judging a person's moral character (Rusconi et al., 2020; Skowronski & Carlston, 1987). Moral behaviors, such as helping a stranger who has fallen, are perceived to be relatively undiagnostic of moral character, because both moral and immoral people are assumed to perform such socially expected behaviors. Immoral behaviors, such as stealing from a colleague, on the other hand, are perceived to be highly diagnostic of immoral character, as only immoral people are assumed to perform them (Skowronski & Carlston, 1987). Consequently, a person who both helped a stranger and stole from a colleague would likely be perceived as immoral overall (Rusconi et al., 2020).

Similarly, many political opinions or other behavioral cues may be highly diagnostic of political categories or may at least be perceived as such. Especially for politicians, who in most cases adhere to the party line, opinions should be a highly diagnostic indicator of their party, assuming sufficient ideological differences between parties. For the mass public, on the other hand, a different picture emerges. As I argued above, levels of ideological divergence and alignment are rather low with at best modest increases over time. In contrast, perceptions of ideological divergence and group homogeneity are sometimes dramatically exaggerated. This suggests that high perceived rather than high actual diagnosticity may drive categorizations and their robustness. In Chapter 3 we investigated whether it is possible to reduce the perceived diagnosticity of political opinions. Across three experiments we showed that counterstereotypes designed to reduce the diagnosticity of political opinions can decrease political categorization on the basis of these opinions. For this project we focused on broad ideological and party groups in the US (Conservatives and Liberals, Republicans and Democrats).

Third Article

Just because we infer category memberships does not automatically mean that we perceive the targets as instances of the respective categories rather than individuals. Categorical information exists along a continuum from loosely connected individuals to cohesive social groups. Consider the difference between introverted people and members of a sports team. While introversion can be seen as defining a category, introverted people are rarely perceived as a unified group—they appear instead as scattered individuals who share a single trait but little else. A sports team, by contrast, registers as highly cohesive: members pursue common goals, share overlapping characteristics, and maintain strong internal connections. This perceived unity and coherence is called entitativity. Political categories likely exhibit similar variation in entitativity. For instance, “Homophobes” might be perceived as a low-entitativity category—individuals who share an attitude but lack organizational structure or collective identity. Groups like Antifa, however, are likely perceived as high-entitativity units with coordinated action and shared purpose.

Importantly, entitativity perceptions may not only vary between categories but also over time. It seems plausible that in polarizing societies, people may come to perceive political categories as increasingly entitative. Such a shift would matter because entitativity has been shown to increase a range of intergroup biases (Abelson et al., 1998; Callahan & Ledgerwood, 2016; Castano et al., 2003; Cikara et al., 2014; Crawford et al., 2002; Dang et al., 2018; Moscatelli & Rubini, 2013; Newheiser et al., 2012; Rogier & Yzerbyt, 1999; Rydell et al., 2007; Sjöström & Gollwitzer, 2015; Spencer-Rodgers et al., 2007; Yzerbyt et al., 1998). In essence, high entitativity transforms categories into groups and individuals into interchangeable group members. This is starkly illustrated by research on displaced revenge, showing that entitativity increases the satisfaction people derive from retaliating against *any* group member for harm caused by another (Sjöström & Gollwitzer, 2015). Given the important role of entitativity in determining the effects of political categorization, it became the focus of the third dissertation article.

One way through which the entitativity perceptions of political groups may change over time is through language-based communication. The use of linguistic group labels (like “Republicans” or “Democrats”) is argued to be both an expression and a driver of high perceived entitativity (Beukeboom & Burgers, 2019). By using group labels frequently and in specific ways, perceptions of high entitativity can thus be culturally transmitted. We investigated this phenomenon on social media, where labeling is ubiquitous. Using fine-tuned transformer models to classify Facebook posts and comments, we found that US party labels (e.g., “Republicans” or “Democrats”) reliably predict higher levels of partisan prejudice in comment sections.

Together, these three articles examine the process of political categorization from its spontaneous occurrence in everyday contexts (Chapter 2), through the role of perceived diagnosticity in driving categorical inferences (Chapter 3), to how linguistic practices shape the entitativity of political groups and associated prejudice (Chapter 4). Each article is presented in full in the following chapters.

References

Abelson, R. P., Dasgupta, N., Park, J., & Banaji, M. R. (1998). Perceptions of the collective other. *Personality and Social Psychology Review*, *2*(4), 243–250. https://doi.org/10.1207/s15327957pspr0204_2

Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, *70*(2), 542–555. https://doi.org/10.1017/S0022381608080493

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *Journal of Politics*, *76*(3), 607–620. https://doi.org/10.1017/S0022381614000085

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, *80*(3), 964–981. https://doi.org/10.1086/697253

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, *114*(2), 408–446. https://doi.org/10.1086/590649

Beukeboom, C. J., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, *7*, 1–37. https://doi.org/10.12840/issn.2255-4165.017

Borbáth, E., Hutter, S., & Leininger, A. (2023). Cleavage politics, polarisation and participation in Western Europe. *West European Politics*, *46*(4), 631–651. https://doi.org/10.1080/01402382.2022.2161786

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2024). Cross-country trends in affective polarization. *Review of Economics and Statistics*, *106*(2), 557–565. https://doi.org/10.1162/rest_a_01160

Boykin, K., Brown, M., Macchione, A. L., Drea, K. M., & Sacco, D. F. (2021). Noncompliance with masking as a coalitional signal to US conservatives in a pandemic. *Evolutionary Psychological Science*, *7*(3), 232–238. https://doi.org/10.1007/s40806-021-00277-x

Broockman, D. E., Kalla, J. L., & Westwood, S. J. (2023). Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science*, *67*(3), 808–828. https://doi.org/10.1111/ajps.12719

Bunosso, I., Donnelly, G. E., & Malkoc, S. A. (2022). Is this store liberal (un)like me?: Consumer response to COVID-19 policies in a politically polarized world. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4106797

Caldwell, D., Cohen, G., & Vivyan, N. (2025). Long-run trends in partisan polarization of climate policy-relevant attitudes across countries. *Environmental Politics*, *34*(5), 767–792. https://doi.org/10.1080/09644016.2024.2403957

Callahan, S. P., & Ledgerwood, A. (2016). On the psychological function of flags and logos: Group identity symbols increase perceived entitativity. *Journal of Personality and Social Psychology*, *110*(4), 528–550. https://doi.org/10.1037/pspi0000047

Carlin, R. E., & Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior*, *35*(1), 43–63. https://doi.org/10.1007/s11109-011-9181-x

Castano, E., Sacchi, S., & Gries, P. H. (2003). The perception of the other in international relations: Evidence for the polarizing effect of entitativity. *Political Psychology*, *24*(3), 449–468. https://doi.org/10.1111/0162-895X.00336

Cely, T. (2025). One more constrained than the other: Asymmetrical ideological alignment and its implications for polarization. *European Journal of Political Research*, *64*(4), 1945–1973. https://doi.org/10.1111/1475-6765.70016

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125. https://doi.org/10.1016/j.jesp.2014.06.007

Colombatto, C., Uddenberg, S., & Scholl, B. J. (2021). The efficiency of demography in face perception. *Attention, Perception, & Psychophysics*, *83*(8), 3104–3117. https://doi.org/10.3758/s13414-021-02351-9

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology*, *83*(5), 1076–1094. https://doi.org/10.1037/0022-3514.83.5.1076

Dang, J., Liu, L., Ren, D., & Gu, Z. (2018). "Groupy" allies are more beneficial while "groupy" enemies are more harmful. *Social Psychological and Personality Science*, *9*(8), 925–934. https://doi.org/10.1177/1948550617729409

Davies, M. (2016). *Corpus of news on the web (NOW)* [Dataset]. https://www.english-corpora.org/now/

Dias, N. C., & Lelkes, Y. (2022). The nature of affective polarization: Disentangling policy disagreement from partisan identity. *American Journal of Political Science*, *66*(3), 775–790. https://doi.org/10.1111/ajps.12628

Dias, N. C., Lelkes, Y., & Pearl, J. (2025). American partisans vastly under-estimate the diversity of other partisans' policy attitudes. *Political Science Research and Methods*, *13*(3), 725–735. https://doi.org/10.1017/psrm.2024.36

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2022). (Mis)estimating affective polarization. *The Journal of Politics*, *84*(2), 1106–1117. https://doi.org/10.1086/715603

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences*, *22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Enders, A. M. (2021). Issues versus affect: How do elite and mass polarization compare? *The Journal of Politics*, *83*(4), 1872–1877. https://doi.org/10.1086/715059

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science*, *11*(1), 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Garzia, D., Ferreira Da Silva, F., & Maye, S. (2023). Affective polarization in comparative and longitudinal perspective. *Public Opinion Quarterly*, *87*(1), 219–231. https://doi.org/10.1093/poq/nfad004

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, *15*(1), 183–217. https://doi.org/10.1080/10463280440000026

Gestefeld, M., Lorenz, J., Henschel, N. T., & Boehnke, K. (2022). Decomposing attitude distributions to characterize attitude polarization in Europe. *SN Social Sciences*, *2*(7), Article 110. https://doi.org/10.1007/s43545-022-00342-7

Gift, K., & Gift, T. (2015). Does politics influence hiring? Evidence from a randomized experiment. *Political Behavior*, *37*(3), 653–675. https://doi.org/10.1007/s11109-014-9286-0

Gimpel, J. G., Newton, J., & Reeves, A. (2025). Short-haul moves and the political geography of partisanship: Intrametropolitan migration as a force for change in U.S. politics. *Urban Affairs Review*. Advance online publication. https://doi.org/10.1177/10780874251380677

Gummer, T., & Oehrlein, A.-S. (2025). Using Google trends data to study high-frequency search terms: Evidence for a reliability-frequency continuum. *Social Science Computer Review*, *43*(4), 814–826. https://doi.org/10.1177/08944393241279421

Hare, C. (2022). Constrained citizens? Ideological structure and conflict extension in the US electorate, 1980–2016. *British Journal of Political Science*, *52*(4), 1602–1621. https://doi.org/10.1017/S000712342100051X

Harteveld, E., & Wagner, M. (2023). Does affective polarisation increase turnout? Evidence from Germany, The Netherlands and Spain. *West European Politics*, *46*(4), 732–759. https://doi.org/10.1080/01402382.2022.2087395

Hartman, R., Blakey, W., Womick, J., Bail, C., Finkel, E. J., Han, H., Sarrouf, J., Schroeder, J., Sheeran, P., Van Bavel, J. J., Willer, R., & Gray, K. (2022). Interventions to reduce partisan animosity. *Nature Human Behaviour*, *6*(9), 1194–1205. https://doi.org/10.1038/s41562-022-01442-3

Hebenstreit, J. (2023). Voter polarisation in Germany: Unpolarised western but polarised eastern Germany? *German Politics*, *32*(1), 63–84. https://doi.org/10.1080/09644008.2022.2056595

Huber, G. A., & Malhotra, N. (2017). Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics*, *79*(1), 269–283. https://doi.org/10.1086/687533

Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological Science*, *7*(5), 496–503. https://doi.org/10.1177/1745691612448792

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Janssen, L., & Turkenburg, E. (2025). Breaking free from linear assumptions: Unravelling the relationship between affective polarization and democratic support. *European Journal of Political Research*, *64*(3), 1465–1479. https://doi.org/10.1111/1475-6765.12725

Jones, J., & Cisternino, I. (2024). LGBTQ visibility measured consistently and persistently in US Twitter bios from 2012 through 2023. *Journal of Quantitative Description: Digital Media*, *4*. https://doi.org/10.51685/jqd.2024.017

Kozlowski, A. C., & Murphy, J. P. (2021). Issue alignment and partisanship in the American public: Revisiting the 'partisans without constraint' thesis. *Social Science Research*, *94*, Article 102498. https://doi.org/10.1016/j.ssresearch.2020.102498

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, *4*(3), 279–286. https://doi.org/10.1038/s41562-019-0766-4

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, *80*(S1), 392–410. https://doi.org/10.1093/poq/nfw005

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly*, *80*(S1), 378–391. https://doi.org/10.1093/poq/nfv045

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, *51*(1), 93–120. https://doi.org/10.1146/annurev.psych.51.1.93

Mandalaywala, T. M., & Legaspi, J. K. (2023). Automatic encoding across social categories in American children and adults. *Developmental Psychology*, *59*(12), 2296–2303. https://doi.org/10.1037/dev0001578

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*, *117*(26), 14864–14872. https://doi.org/10.1073/pnas.2001263117

Moscatelli, S., & Rubini, M. (2013). The impact of group entitativity on negative outcome allocations. *The Journal of Social Psychology*, *153*(2), 149–160. https://doi.org/10.1080/00224545.2012.712068

Myers, C. D., & Hvidsten, T. (2024). Who is the party? The effect of counter-stereotypical partisan exemplars on inter-party stereotypes and affect. *Political Behavior*. https://doi.org/10.1007/s11109-024-09991-1

Nelson, M. H. (2022). Resentment is like drinking poison? The heterogeneous health effects of affective polarization. *Journal of Health and Social Behavior*, *63*(4), 508–524. https://doi.org/10.1177/00221465221075311

Newheiser, A.-K., Sawaoka, T., & Dovidio, J. F. (2012). Why do we punish groups? High entitativity promotes moral suspicion. *Journal of Experimental Social Psychology*, *48*(4), 931–936. https://doi.org/10.1016/j.jesp.2012.02.013

Ojer, J., Cárcamo, D., Pastor-Satorras, R., & Starnini, M. (2025). Charting multidimensional ideological polarization across demographic groups in the USA. *Nature Human Behaviour*, *9*(10), 2027–2037. https://doi.org/10.1038/s41562-025-02251-0

Orhan, Y. E. (2022). The relationship between affective polarization and democratic backsliding: Comparative evidence. *Democratization*, *29*(4), 714–735. https://doi.org/10.1080/13510347.2021.2008912

Orr, L. V., & Huber, G. A. (2020). The policy basis of measured partisan animosity in the United States. *American Journal of Political Science*, *64*(3), 569–586. https://doi.org/10.1111/ajps.12498

Park, B. (2018). How are we apart? Continuity and change in the structure of ideological disagreement in the American public, 1980–2012. *Social Forces*, *96*(4), 1757–1784. https://doi.org/10.1093/sf/sox093

Phills, C. E., Hahn, A., & Gawronski, B. (2020). The bidirectional causal relation between implicit stereotypes and implicit prejudice. *Personality and Social Psychology Bulletin*, *46*(9), 1318–1330. https://doi.org/10.1177/0146167219899234

Piazza, J. A. (2023). Political polarization and political violence. *Security Studies*, *32*(3), 476–504. https://doi.org/10.1080/09636412.2023.2225780

Powdthavee, N., Riyanto, Y. E., Wong, E. C. L., Yeo, J. X. W., & Chan, Q. Y. (2021). When face masks signal social identity: Explaining the deep face-mask divide during the COVID-19 pandemic. *PLOS ONE*, *16*(6), Article e0253195. https://doi.org/10.1371/journal.pone.0253195

Rogers, N., & Jones, J. J. (2021). Using Twitter bios to measure changes in self-identity: Are Americans defining themselves more politically over time? *Journal of Social Computing*, *2*(1), 1–13. https://doi.org/10.23919/JSC.2021.0002

Rogier, A., & Yzerbyt, V. (1999). Social attribution, correspondence bias, and the emergence of stereotypes. *Swiss Journal of Psychology / Schweizerische Zeitschrift Für Psychologie / Revue Suisse de Psychologie*, *58*(4), 233–240. https://doi.org/10.1024/1421-0185.58.4.233

Roth, J., Deutsch, R., & Sherman, J. W. (2019). Automatic antecedents of discrimination. *European Psychologist*, *24*(3), 219–230. https://doi.org/10.1027/1016-9040/a000321

Rothers, A. (2025). The poles in polarization: Social categorization and affective polarization in multiparty systems. *Electoral Studies*, *95*, Article 102908. https://doi.org/10.1016/j.electstud.2025.102908

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., Bursalıoğlu, A., Bühler, V., Čadek, M., Çetinçelik, M., Clay, G., Cortijos-Bernabeu, A., Damnjanović, K., Dugue, T. M., Esberg, M., Esteban-Serna, C., … Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour*, *5*(10), 1369–1380. https://doi.org/10.1038/s41562-021-01092-x

Rusconi, P., Sacchi, S., Brambilla, M., Capellini, R., & Cherubini, P. (2020). Being honest and acting consistently: Boundary conditions of the negativity effect in the attribution of morality. *Social Cognition*, *38*(2), 146–178. https://doi.org/10.1521/soco.2020.38.2.146

Rydell, R. J., Hugenberg, K., Ray, D., & Mackie, D. M. (2007). Implicit theories about groups and stereotyping: The role of group entitativity. *Personality and Social Psychology Bulletin*, *33*(4), 549–558. https://doi.org/10.1177/0146167206296956

Shafranek, R. M. (2021). Political considerations in nonpolitical decisions: A conjoint analysis of roommate choice. *Political Behavior*, *43*(1), 271–300. https://doi.org/10.1007/s11109-019-09554-9

Simas, E. N., Clifford, S., & Kirkland, J. H. (2020). How empathic concern fuels political polarization. *American Political Science Review*, *114*(1), 258–269. https://doi.org/10.1017/S0003055419000534

Sjöström, A., & Gollwitzer, M. (2015). Displaced revenge: Can revenge taste "sweet" if it aims at a different target? *Journal of Experimental Social Psychology*, *56*, 191–202. https://doi.org/10.1016/j.jesp.2014.09.016

Skowronski, J. J. (2002). Honesty and intelligence judgments of individuals and groups: The effects of entity-related behavior diagnosticity and implicit theories. *Social Cognition*, *20*(2), 136–169. https://doi.org/10.1521/soco.20.2.136.20993

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*, 689–699. https://doi.org/10.1037/0022-3514.52.4.689

Spencer-Rodgers, J., Hamilton, D. L., & Sherman, S. J. (2007). The central role of entitativity in stereotypes of social categories and task groups. *Journal of Personality and Social Psychology*, *92*(3), 369–388. https://doi.org/10.1037/0022-3514.92.3.369

Tajfel, H., & Turner, J. C. (1979). An integrative theory of inter-group conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of inter-group relations* (pp. 33–47). Brooks/Cole.

Torcal, M., & Thomson, Z. A. (2023). Social trust and affective polarization in Spain (2014–19). *Electoral Studies*, *81*, Article 102582. https://doi.org/10.1016/j.electstud.2023.102582

Tucker, L., & Jones, J. (2026). Pronoun lists in profile bios display increased prevalence, systematic co-presence with other keywords and network tie clustering among US Twitter users 2015-2022. *Journal of Quantitative Description: Digital Media*, *3*. https://doi.org/10.51685/jqd.2023.003

Voelkel, J. G., Stagnaro, M. N., Chu, J. Y., Pink, S. L., Mernyk, J. S., Redekopp, C., Ghezae, I., Cashman, M., Adjodah, D., Allen, L. G., Allis, L. V., Baleria, G., Ballantyne, N., Van Bavel, J. J., Blunden, H., Braley, A., Bryan, C. J., Celniker, J. B., Cikara, M., … Willer, R. (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, *386*(6719), Article eadh4764. https://doi.org/10.1126/science.adh4764

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, *69*, Article 102199. https://doi.org/10.1016/j.electstud.2020.102199

Warncke, P. (2025). What explains country-level differences in political belief system coherence? *Political Behavior*, *47*(4), 1853–1876. https://doi.org/10.1007/s11109-025-10015-9

Webster, S. W., & Abramowitz, A. I. (2017). The ideological foundations of affective polarization in the U.S. electorate. *American Politics Research*, *45*(4), 621–647. https://doi.org/10.1177/1532673X17703132

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, *10*(2), 145–158. https://doi.org/10.1177/1745691615569849

Wissenschaft im Dialog. (2025). *Wissenschaftsbarometer 2025*. https://wissenschaft-im-dialog.de/projekte/wissenschaftsbarometer

Yudkin, D. A., Hawkins, S., & Dixon, T. (2019). *The perception gap: How false impressions are pulling americans apart*. More in Common. https://doi.org/10.31234/osf.io/r3h5q

Yzerbyt, V. Y., Rogier, A., & Fiske, S. T. (1998). Group entitativity and social attribution: On translating situational constraints into stereotypes. *Personality and Social Psychology Bulletin*, *24*(10), 1089–1103. https://doi.org/10.1177/01461672982410006

CHAPTER 2

Investigating Mechanisms of Political Polarization: Perceivers Spontaneously Infer Ideological Categories from Other People's Behavior

Sander, C. W., & Degner, J. (2024). Investigating mechanisms of political polarization: Perceivers spontaneously infer ideological categories from other people's behavior. *Journal of Applied Social Psychology*, *54*(1), 5–17. https://doi.org/10.1111/jasp.13011

Check for updates

Received: 16 June 2023 | Revised: 24 August 2023 | Accepted: 28 August 2023

DOI: 10.1111/jasp.13011

ORIGINAL ARTICLE

Journal of Applied Social Psychology WILEY

Investigating mechanisms of political polarization: Perceivers spontaneously infer ideological categories from other people's behavior

Carsten W. Sander | **Juliane Degner**

Department of Social Psychology, Universität Hamburg, Hamburg, Germany

Correspondence
Carsten W. Sander, Department of Social Psychology, Universität Hamburg, Von-Melle-Park 5, Raum 3025a, 20146 Hamburg, Germany.
Email: carsten.sander@uni-hamburg.de

Funding information
This research was supported by a doctoral scholarship of the Studienstiftung des Deutschen Volkes awarded to Carsten W. Sander.

Abstract

When observing others' behavior, people routinely infer personality traits from it. Research on the correspondence bias has shown that they draw these inferences even from behavior that is situationally constrained. Moreover, these inferences often happen spontaneously, that is, when people have no intention of forming an impression. The current research investigated whether the same applies for inferences of ideological categories, such as conservative, feminist, or climate change denier. We present a series of five preregistered experiments ($N = 1012$) employing both direct and indirect measures from the impression formation literature. In Experiments 1 and 2, we found that people deliberately, as well as spontaneously, inferred ideological categories from ideology-implying behavioral descriptions. In Experiments 3, 4, and 5, we found these inferences to be very robust, in that they persist even when alternative reasons for the ideology-implying behaviors are presented, suggesting that they are subject to the correspondence bias. We argue that to the extent that people attribute negative valence (affective polarization) or extreme attitudes (perceived issue polarization) to ideological categories, spontaneous inferences of these categories can serve as a precondition for a range of negative interpersonal outcomes related to political polarization.

1 | INTRODUCTION

Many nations across the globe have experienced political polarization in recent years (Carothers & O'Donohue, 2019; Somer & McCoy, 2018), including Austria, Colombia, Egypt, France, Hungary, Poland, Thailand, Turkey, the United Kingdom, the United States, and Venezuela (McCoy et al., 2018). Although under the right conditions, polarization may have positive long-term consequences for democratization (LeBas, 2018), the overall correlation between levels of polarization and liberal democracy ratings is negative (Somer et al., 2021). Understanding the underlying psychological mechanisms of political polarization thus appears a valuable endeavor. The current research applies social-cognitive research from the domain of person perception to the study of political polarization and investigates a cognitive process and attribution effect that may contribute to polarization's negative consequences.

2 | POLITICAL POLARIZATION

Scholars have broken down the complex concept of political polarization into multiple concrete components. On the one hand, there are components related to the distribution of actual issue

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2023 The Authors. *Journal of Applied Social Psychology* published by Wiley Periodicals LLC.

positions in society: People's political opinions, beliefs, or values can be conceptualized as distributions along multiple dimensions (e.g., pro-choice vs. pro-life, government control vs. free market, and so on). Polarization is characterized by these distributions being bimodal and dispersed—such that there are two groups of people with vastly different opinions (ideological or issue polarization; Fiorina & Abrams, 2008). Moreover, polarization is characterized by different ideological dimensions and other social cleavages being correlated with each other (partisan alignment) and with membership to political groups or parties (partisan sorting; Fiorina & Abrams, 2008; Mason & Wronski, 2018; Park, 2018).

On the other hand, there are components referring to how members of different political groups perceive each other. First, there is affective polarization (Iyengar et al., 2012, 2019), which describes the extent to which partisans dislike and distrust those from the other party. Second, there is perceived issue polarization (Enders & Armaly, 2019), which refers to perceiving out-partisans to have vastly different opinions than oneself—including very negative opinions about one's party (negative meta-perceptions; Fernbach & Van Boven, 2022). Perceived issue polarization is often larger than actual issue polarization (Ahler, 2014; Levendusky & Malhotra, 2016), a phenomenon termed "false polarization" (e.g., Fernbach & Van Boven, 2022). These perception-related components of political polarization have been shown to cause negative interpersonal outcomes such as people being uncomfortable with the idea of talking about politics with an out-partisan and of having out-partisan friends or children-in-law (Broockman et al., 2020; Koudenburg & Kashima, 2022). Thus, to the extent that people hold negative stereotypes and harbor feelings of dislike and distrust towards their out-party, categorizing someone as an out-partisan can be expected to elicit negative interpersonal outcomes.

This line of thought can be applied not only to party categories. To make sense of others' opinions, beliefs, and behaviors, people also employ other ideological categories such as leftist, racist, environmentalist, homophobe, climate change denier, or trans activist. Note that by using the term "ideological categories", we are not suggesting that these categories necessarily describe actual ideologies defined as belief systems or refer to the left-right or liberal-conservative dimensions of politics. As we focus on how groups and group affiliations are perceived, we instead use the term "ideological" as a lay perceiver might do, that is, to describe categories that are perceived as reflecting shared socio-political worldviews, behaviors, and ideologies. These nonparty categories may also be negatively stereotyped, disliked, and distrusted, and perception of category membership may also lead to negative interpersonal outcomes. We therefore think it is important to investigate how people come to perceive others as members of these different kinds of ideological categories. By leveraging research from the domain of impression formation we put forward spontaneous behavior-based ideological inferences as one mechanism by which this categorization could occur in everyday life and investigate how robust these inferences are.

3 | SPONTANEOUS IDEOLOGICAL INFERENCES AND THE CORRESPONDENCE BIAS

Although a person's ideological category membership is most easily recognized from self-disclosing statements (e.g., someone stating, "I am a Democrat") or unambiguous symbols (e.g., a "Vote Republican" bumper sticker), such self-evident indicators are often not available. People might nevertheless easily infer a person's ideological category. Research in person perception has shown that people tend to infer others' desires, beliefs, and personality traits from ambiguous external cues (e.g., demographics, clothing style, social class attributes, and so on) or from their behavior (e.g., Malle & Holbrook, 2012; Uleman et al., 2012). They sometimes do so even when it makes little sense. For instance, in a seminal experiment by Jones and Harris (1967), participants were presented with an essay expressing either a pro-Castro or an anti-Castro attitude, ostensibly written by another student. When asked to rate the student's true attitude towards Castro, participants unsurprisingly answered in line with the attitude expressed in the essay. However, they drew the same conclusion even when they had been told that the student did not choose sides freely but had been assigned a side. Apparently, they did not appreciate this strong situational constraint when judging the essay-writer. This tendency to draw dispositional inferences even when behavior is situationally constrained is called the correspondence bias and has been replicated numerous times and across a broad range of paradigms and dispositions (for reviews, see Gawronski, 2004; Gilbert & Malone, 1995).

Some researchers (e.g., Gilbert & Malone, 1995), defined the correspondence bias in a strictly normative manner, namely as the tendency of participants to make dispositional inferences "even if a logical analysis suggests they should not" (p. 21). Other authors (e.g., Gawronski, 2004) place less emphasis on normative adequacy and define the correspondence bias as the tendency to draw dispositional inferences "even when the observed behavior was highly constrained by situational factors" (p. 183). We follow the latter definition, because when it comes to social categorizations, it is difficult to determine what logically should be inferred. For example, in the case of Jones & Harris's (1967) manipulation, it is not clear how strong the situational constraint was, as we lack information on how many people would comply with the assignment of writing a counter-attitudinal essay on that specific topic. Therefore, from our perspective, the term correspondence bias does not necessarily imply that a normatively incorrect inference was drawn. Rather we use the term to denote the robustness of the inference against plausible alternative reasons for the behavior.

According to classic theoretical models of social inference, the correspondence bias occurs, because drawing dispositional inferences is a relatively automatic process, whereas the situational correction of these inferences requires motivated and effortful processing (Gilbert, 1989). In other words, although we reliably and effortlessly draw dispositional inferences from others' behavior, we might only correct these inferences occasionally, namely when we

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

have enough motivation and cognitive resources (Gawronski, 2004). The automatic nature of initial dispositional inference was demonstrated by researchers who asked participants to merely read or memorize behavioral descriptions (rather than to give explicit trait judgments) and then applied memory or other cognitive tasks to indirectly assess inferences. They found that so-called spontaneous trait inferences (STIs) usually occur without the perceiver's intention (Todorov & Uleman, 2002), are reduced but not eliminated by cognitive load (Wells et al., 2011), and even occur when participants are instructed to suppress them (Shimizu, 2017). Moreover, a recent meta-analysis suggests that STI effects are very robust (Bott et al., 2022). The correspondence bias thus might occur, because STIs are relatively automatic while their correction is a more controlled and effortful process (Moskowitz, 2005).

Albeit ideological and trait categories may differ—with ideological categories likely being perceived as less stable and more controllable than trait categories—we nevertheless expect that these mechanisms also apply to ideological inferences: Similar to traits, we may infer partisanship as well as other kinds of ideological categories from other's behavior. For example, we may infer from someone saying that they oppose gay marriage that they are a conservative or from someone posting pictures from a Women's March on social media that they are a feminist. Just as trait inferences, ideological inferences may occur spontaneously, that is, independent of whether or not people have the intention of forming an impression. And just as trait inferences, they may be subject to the correspondence bias. That is, ideological inferences may subsist even when observers learn that the underlying behavior was situationally constrained. To the extent that people arrive at rigid ideological categorizations via spontaneous ideological inferences (SIIs), these inferences may act as a precursor to negative interpersonal outcomes of affective and perceived issue polarization. Although previous research has shown that some ideological categories such as party affiliations can be inferred from individuals' political opinions (Lüders et al., 2023), no research so far investigated whether such inferences also occur spontaneously. However, spontaneity is essential for real-life social impact. If inferences only occur when prompted and when people engage in controlled and deliberate thought and decision processes, inferences and their assumed negative interpersonal outcomes may rarely occur in real life. Thus, it is crucial to investigate whether such inferences do occur spontaneously. This can be achieved by using indirect measures, which do not force participants to make inferences.

4 | THE CURRENT RESEARCH

The current research investigates whether people infer a broad range of ideological categories from other people's behavior and to what extent the correspondence bias occurs for these inferences. As different experimental procedures might elicit different processes, we pursue a multimethod approach using both direct and indirect measures. For the direct measure, we simply present participants with ideology-implying behavioral descriptions (e.g., "Sarah has an appointment today at the Citizens Registration Office to register that she is leaving the church") and explicitly ask them to make judgments about the actors (e.g., "Do you think Sarah is an atheist?"). As this should motivate them to form impressions and devote cognitive capacity to the task, this measure should reflect intentional and deliberate impression formation. In the indirect paradigms borrowed from the STI literature, participants are instructed to merely memorize behavioral descriptions rather than to make dispositional judgments. In a subsequent indirect test phase, they are then presented with the previously implied ideological category (e.g., "atheist"; implied condition) or a category implied in a different behavioral description serving as a control (e.g., "esoteric"; implied-other condition). Participants must indicate via keypress whether the presented ideological category appeared in the behavioral description or not. If participants have inferred the implied category, this should make it more difficult to indicate that the category did not appear in the statement. Accordingly, longer response latencies for correct rejections or higher rates of false recognitions in the implied compared with the implied-other condition can be interpreted as evidence that the implied category was inferred (e.g., Todd et al., 2011). As under these conditions participants should have no intention of forming an impression and may even be unaware of doing so, the indirect measures should predominantly reflect spontaneous processing.

In Experiment 1, we use a direct measure to test whether participants infer ideological categories when explicitly asked about their impressions. In Experiment 2, we use the indirect probe recognition paradigm (e.g., Todd et al., 2011) to test whether they also do so spontaneously. In Experiment 3, we go back to a direct measure to test whether ideological inferences are reduced when plausible alternative reasons for the behaviors are presented. In Experiments 4 and 5, we again use the probe recognition paradigm as well as another indirect paradigm—the false recognition paradigm (Todorov & Uleman, 2002)—to test whether the correspondence bias occurs for spontaneous inferences. We recruited participants for all five experiments via the online platform Prolific (www.prolific.co), inviting only users who reported to be of German nationality, native speakers of German, and currently located in Germany. All experiments were run using Qualtrics (www.qualtrics.com) and PsychoPy (Peirce et al., 2019). Analyses were performed using R (R Core Team, 2022) and the rstatix package (Kassambara, 2023). We report sample size determination, data exclusions, manipulations, and all measures in the description of each experiment and in the Supporting Information. All procedures were approved by the faculty's local ethics committee (protocol number: 2022_006).

5 | EXPERIMENT 1

In the first experiment, we used a direct measure to investigate whether ideological categories are inferred when participants are explicitly asked about their impressions. We constructed the direct measure in such a way that it was structurally very similar to the indirect measure we employed in Experiment 2.

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.1 | Method

5.1.1 | Sample

We collected valid data from 27 participants. This sample size was predetermined using G*Power 3.1 (Faul et al., 2007) to obtain 80% power for detecting a medium effect of $d_z = 0.50$ at $\alpha = .05$ using a one-tailed dependent samples t test. For demographic data for all experiments see Table 1 (for additional data regarding political orientation, political interest, and political satisfaction see Supporting Information: Table S4.1). Participants were compensated 1.35 GBP for completing the 9 min experiment.

5.1.2 | Materials

We developed behavioral statements that imply ideological categories without explicitly naming them (e.g., the behavioral statement "When Michelle learned that there was no soy milk, she decided not to buy coffee after all." was judged to imply the category "vegan"). Details of stimulus generation and extensive pretesting are provided in the Supporting Information: S1. For Experiment 1, we selected a set of 24 statements (stimulus set 1; see Supporting Information: Table S1.1). Each statement implied an ideological category (implied condition) and was assigned a semantically distinct category implied by another statement (implied-other condition), which was chosen such that its word length was close to the respective implied category (important for Experiment 2).

5.1.3 | Procedure

Participants were informed that the experiment was about text comprehension and instructed to read and memorize a series of short statements. They were presented 24 ideology-implying target statements in a randomized order. The duration of presentation was based on statement length (300 ms per word) and ranged from 2.7 to 12.3 s. After a 250 ms blank screen following each statement, participants were asked whether they thought the targets belonged to two ideological categories, using a 5-point Likert-like scale ranging from 1 = "*no*" to 5 = "*yes*." One was the category implied by the statement (implied condition). The other was a semantically distinct category implied by another target statement (implied-other condition) that served as a control. This procedure is unusual, because with direct measures there is usually no need for a control condition. We implemented it to ensure that the effect size calculation was comparable to the indirect measure used in Experiment 2, for which such a control condition is necessary. Participants were told that there are no right or wrong answers and instructed to respond spontaneously. Participants first performed a practice trial, followed by the 24 test trials. Details on further measures and on debriefing for all five experiments can be found in the Supporting Information: S3 and S4.

TABLE 1 Demographic data for all samples.

Variable	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Gender					
Female	14	84	88	126	183
Male	13	76	109	101	196
Other	0	3	2	5	6
Undisclosed	0	0	0	4	1
Age					
M (SD)	31.4 (8.7)	29.9 (10.4)	27.1 (8.0)	27.8 (7.9)	27.5 (7.7)
Range	18–54	18–69	18–62	18–66	18–67

5.2 | Results and discussion

We computed the mean rating of implied and implied-other categories for each participant. The preregistered one-tailed dependent samples t test revealed a significant effect of probe type, $t(26) = 27.4$, $p < .001$, $d_z = 5.27$. On average, participants judged the targets to belong to the implied categories ($M = 4.03$, SD = 0.42) to a significantly stronger extent than to the implied-other categories ($M = 1.70$, SD = 0.34). We conclude from this that participants did indeed infer the ideological categories that were implied in the statements.

The results of Experiment 1 provide first evidence that people infer a broad range of ideological categories from other people's behavior. This experiment used direct measures in which participants were explicitly asked to form impressions of the target persons. As we have argued, direct measures should reflect rather deliberate processing because they might increase participants' motivation and cognitive capacity devoted to forming impressions. Indirect measures should not have this effect, because they do not rely on participants' intention to form impressions. These measures should therefore predominantly reflect spontaneous processing. Moreover, indirect measures may be less likely than direct measures to be affected by socially desirable responding and demand effects (Corneille & Lush, 2023), because they allow for less strategical control over the responses and better conceal the research hypothesis. By using an indirect measure in Experiment 2, we investigate whether ideological inferences also occur spontaneously and simultaneously address these limitations.

6 | EXPERIMENT 2

We used the indirect probe recognition paradigm (e.g., Todd et al., 2011) to investigate whether ideological categories are inferred spontaneously, that is when participants are not explicitly asked about their impressions. As described above, in each trial of this

15598160, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

paradigm participants read a behavioral description and then indicate whether a series of probe words appeared in it or not. The probes contain the implied ideological category, as well as a category implied in another behavioral description (implied-other). Inferring the implied category should make it more difficult to indicate that the respective probe did not appear in the statement. Accordingly, longer response latencies for correct rejections in the implied compared to the implied-other condition can be interpreted as evidence for SIIs.

6.1 | Method

6.1.1 | Sample

Data analyses are based on valid data from $N = 163$ participants. Despite the meta-analytic effect for STIs being moderate to large ($d_z = 0.59$; Bott et al., 2022), we set our smallest effect size of interest to $d_z = 0.20$, due to the novelty of our materials. Using G*Power 3.1 (Faul et al., 2007), we calculated a minimum sample of $N = 156$ to obtain 80% power for detecting this effect at $\alpha = .05$ using a one-tailed dependent samples t test. Following our preregistered exclusion criteria, we excluded five additional participants because their average correct response times were slower than 2 SDs over the sample mean, one who self-reported not having followed the instructions conscientiously, and one who rated their own data to be unfit for analysis. Participants were compensated 3.13 GBP for completing the 25 min experiment.

6.1.2 | Materials

We used the same target statements as in Experiment 1. Each statement was followed by four probes. Target probes were again the implied ideological category (implied condition) and a semantically distinct category implied by another behavioral description (implied-other condition). Both target probes required a negative response (given that they were not presented in the statement). To balance out affirmative and negative responses, we presented two additional filler probes per statement that had actually occurred in the target statement and thus required an affirmative response. To further prevent strategic responding, we included 24 filler statements (see Supporting Information: Table S1.3). These filler statements and their respective probes actually contained the ideological category labels, thus ensuring that the number of required affirmative and negative responses was balanced across all category probes. Moreover, the number of required affirmative and negative responses was balanced across different probe word types (nouns, adjectives, verbs, and names).

6.1.3 | Procedure

Experiment 2 was similar to Experiment 1, except for the following differences. We presented not only target but also filler statements. The order of all statements was randomized. Each statement was followed by a blank screen for 250 ms, a fixation cross for 500 ms, and a randomized sequence of four probes, each of which was presented until a response was made (with a post response pause of 100 ms). Participants were instructed to indicate whether the probe had appeared in the statement or not by pressing [J] for yes or [F] for no, respectively. Responses as well as response latencies were recorded. Participants were required to respond as quickly and accurately as possible. When they made a mistake, an error message appeared (red letter "X" presented for 1000 ms). After presentation of all four probes and an intertrial interval of 500 ms, the next statement was presented. Participants performed a practice block containing 5 statements, after which they received individualized feedback on their error rate. Participants with an error rate higher than 5% were instructed to respond more accurately. They then performed two test blocks of 24 statements each. After completion of the second test block, participants again received summarized feedback on their error rate.

6.2 | Results and discussion

To prepare the response latency data for our main analysis, we applied a cut-off of the individual mean + 2 SDs for slow responses and the log-transformation before aggregating response times for each experimental condition. For reports of descriptive data, we back-transformed the logarithmic mean values by calculating their exponentials. The preregistered one-tailed dependent samples t test revealed a significant effect of probe type, $t(162) = 19.0$, $p < .001$, $d_z = 1.49$. On average, participants responded significantly slower in the implied condition ($M = 799$, SD = 113) than in the implied-other condition ($M = 738$, SD = 94). As there are no conventions regarding outlier correction and transformations, we had preregistered an additional multiverse analysis using different combinations of cutoff criteria and transformations. The effect of probe type was highly significant across all combinations (all $ps < .001$) with effect size estimates ranging from $d_z = 1.21$ to $d_z = 1.72$ (see Supporting Information: Table S5.3).

The results of Experiment 2 provide strong evidence that people spontaneously infer ideological categories when they learn about other people's ideology-implying behavior. Given our use of an indirect paradigm in which participants merely indicated whether certain words had appeared in behavioral descriptions, it is unlikely that participants had the intention of drawing ideological inferences. As actively drawing inferences about the target person would have interfered with this task and slowed down responses, participants may have been rather motivated to suppress such inferences. This suggests that ideological inferences occur with some degree of automaticity. Since the indirect paradigm better conceals the true objective of the experiment, demand effects or socially desirable responding should be less likely than in Experiment 1. Albeit the effect size in Experiment 2 is a lot smaller than in Experiment 1, it is still almost four times as large as the average STI-effect in this

15598816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

paradigm (Bott et al., 2022). This suggests that there may be a higher readiness to infer ideological categories compared to ordinary traits.

In summary, Experiments 1 and 2 provide strong first evidence that people infer ideological categories from other's behavior and that they do so spontaneously, expanding the range of inferences that have been examined in impression formation research to date.

In the following experiments, we investigate the robustness of ideological inferences. Early theorizing on attribution posited the discounting principle, according to which dispositional inferences from a behavior should be reduced when there are plausible alternative reasons for the behavior. Specifically, the principle states that the "role of a given cause in producing a given effect is discounted if other plausible causes are also present" (Kelley, 1973, p. 113). For example, when reading the statement "When Michelle learned that there was no soy milk, she decided not to buy coffee after all," one might infer that Michelle is a vegan. However, knowing that she recently found out that she was lactose intolerant might eliminate that inference—if the discounting principle holds—because the motive of trying to avoid lactose can plausibly explain her behavior. However, as research has demonstrated, the discounting principle often does not hold. People may infer that Michelle is a vegan despite knowing about her lactose intolerance, thus committing the correspondence bias. In the following three experiments, we investigate whether and to what extent ideological inferences are susceptible to discounting information with a direct measure (Exp. 3) and two indirect measures (Exp. 4 and 5).

7 | EXPERIMENT 3

In the third experiment, we went back to using a direct measure to investigate whether ideological inferences subsist when alternative reasons are provided. Experiment 3 was thus identical to Experiment 1 with the exception that we presented the behavioral descriptions together with ancillary statements that either offered alternative reasons of the observed behaviors or not. In selecting reasons, there is a trade-off between making them compelling and realistic. For example, one could possibly explain any behavior by inventing some scenario in which the actor is forced at gunpoint to perform it. However, as situational constraints this extraordinary are unlikely to occur in everyday life, they are uninteresting for real-life application. Accordingly, we have chosen alternative reasons that are probably less potent, but that one might actually encounter in real life.

7.1 | Method

7.1.1 | Sample

As we had no prior information for effect size estimation, we set our smallest effect size of interest to $f = 0.15$ for a two-way interaction effect in a repeated-measures analysis of variance (ANOVA). Using MorePower 6.0.4 (Campbell & Thompson, 2012), we calculated that we would need a sample of $n = 352$ to obtain 80% power for detecting this effect at $\alpha = .05$. However, to avoid unnecessary spending in case of a larger than expected effect size, we employed a sequential testing procedure (Lakens, 2014; Lakens et al., 2021) and preregistered two interim analyses at one and two fourths of the total sample. New alpha boundaries for these analyses were calculated with the rpact package for R (Wassmer & Pahlke, 2020) using the Kim & DeMets alpha spending function (with gammaA set to 1). We also implemented a stopping rule for futility to avoid unnecessary spending in case of an effect size considerably smaller than the smallest effect size of interest with nonbinding futility boundaries calculated using the Kim & DeMets beta spending function (with gammaB set to 1). Due to sequential testing, the total sample size increased to $n = 400$. However, as the effect of interest was significant at the second interim analysis (at $\alpha = .017$), we could stop data collection early at $n = 200$ valid data sets. Following our preregistered exclusion criteria, we excluded three additional participants who rated their data unfit for analysis. Participants were compensated 2.85 GBP for completing the 19 min experiment.

7.1.2 | Materials

For this experiment, we used another set of 24 items, consisting of ideology-implying statements and related ancillary statements for each behavior (stimulus set 2; see Supporting Information: Table S1.2). The ancillary statements either offered sufficient alternative reasons of the observed behaviors (e.g., "Michelle's gastroenterologist recently said her symptoms stemmed from lactose intolerance.") or control statements that consisted of more or less the same words but did not offer an alternative explanation of observed behaviors (e.g., "Michelle's friend was told by his gastroenterologist that his symptoms did not stem from lactose intolerance."). Details of stimulus generation and extensive pretesting are provided in the Supporting Information: S1.

7.1.3 | Procedure

The procedure was similar to Experiment 1. Participants were again presented 24 ideology-implying statements in a randomized order—only this time together with an ancillary statement that either provided a reason for the behavior or not and was presented either directly before or after the behavioral statement. Ancillary statement (reason vs. control) and order of presentation (behavior first vs. ancillary statement first) were manipulated orthogonally within participants using a Latin square design, such that every participant saw six items in each of the four conditions and every item was presented equally often in each condition across all participants. The duration of presentation of both stimuli was again based on statement length (300 ms per word) and ranged from 6.6 to 18.6 s. After rating a target person on the implied and implied-other categories, participants also rated them on trustworthiness and

15598160, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

valence (for analyses of these ratings see Supporting Information: S5.5). Following the main task, participants completed a manipulation check, in which they rated the degree to which each presented ancillary statement provided a sufficient explanation of the respective behavior on a five-point Likert-like scale ranging from 1 = "*no*" to 5 = "*yes*" (this manipulation check was repeated in Experiments 4 and 5).

7.2 | Results and discussion

The manipulation checks of Experiments 3, 4, and 5 revealed that participants rated the ancillary statements containing reasons to explain the behaviors more sufficiently than the ones without reasons (all *p*s < .001, d_z ranging from 1.39 to 1.72; for details see Supporting Information: Table S2.1).

We again calculated mean ratings for each experimental condition and conducted a preregistered two-way repeated-measures ANOVA to evaluate the effects of probe type and ancillary statement. We observed significant main effects for probe type, $F(1, 199) = 1986.70$, $p < .001$, $\eta^2_p = .909$ and ancillary statement, $F(1, 199) = 15.01$, $p < .001$, $\eta^2_p = .07$, as well as a significant interaction, $F(1, 199) = 14.25$, $p < .001$, $\eta^2_p = .067$ (see Figure 1). Posthoc tests confirmed significantly lower ratings of implied probes in the reason compared to the control condition, $t(199) = 6.24$, $p < .001$, $d_z = 0.44$, but not for implied-other probes, $t(199) = 0.525$, $p = .600$, $d_z = 0.04$. However, the effect of probe type remained significant and large for both control statements, $t(199) = 35.9$, $p < .001$, $d_z = 2.54$, and reason statements, $t(199) = 36.1$, $p < .001$, $d_z = 2.56$. We thus conclude that participants again inferred the ideological categories implied in the statements. Although presenting them with alternative reasons for the described behaviors significantly reduced ideological inferences, it did so only to a small degree.

The results of Experiment 3 show that ideological inferences from a behavior can be discounted when reasons for that behavior are presented. However, the discounting effect was very small and did not eliminate the inferences. These results are therefore in line with the assumption of ideological inferences being subject to the correspondence bias. This should not be taken to mean that participants committed logical errors. Rather these results speak to the robustness of the observed ideological inferences against alternative attributions. Apparently, simply providing good reasons for a given behavior does not prevent people from drawing ideological inferences.

As we have argued, asking participants directly for their judgments likely motivated them to form impressions and devote cognitive capacity to the task. These results should therefore reflect more deliberate processing. As deliberate processing may facilitate the effortful situational correction of the initial ideological inferences, the direct measure may have artificially increased the discounting effect compared to what would happen spontaneously in a more naturalistic setting. In Experiment 4, we therefore used an indirect measure—which should be less likely to elicit deliberate processing—and investigated the correspondence bias in SIIs.

8 | EXPERIMENT 4

In the fourth experiment, we combined the indirect paradigm from Experiment 2 with the materials from Experiment 3 to investigate whether and to what extent the correspondence bias occurs under spontaneous processing conditions.

8.1 | Method

8.1.1 | Sample

As we had no prior information for effect size estimation, we again set our smallest effect size of interest of $f = 0.15$. Using MorePower

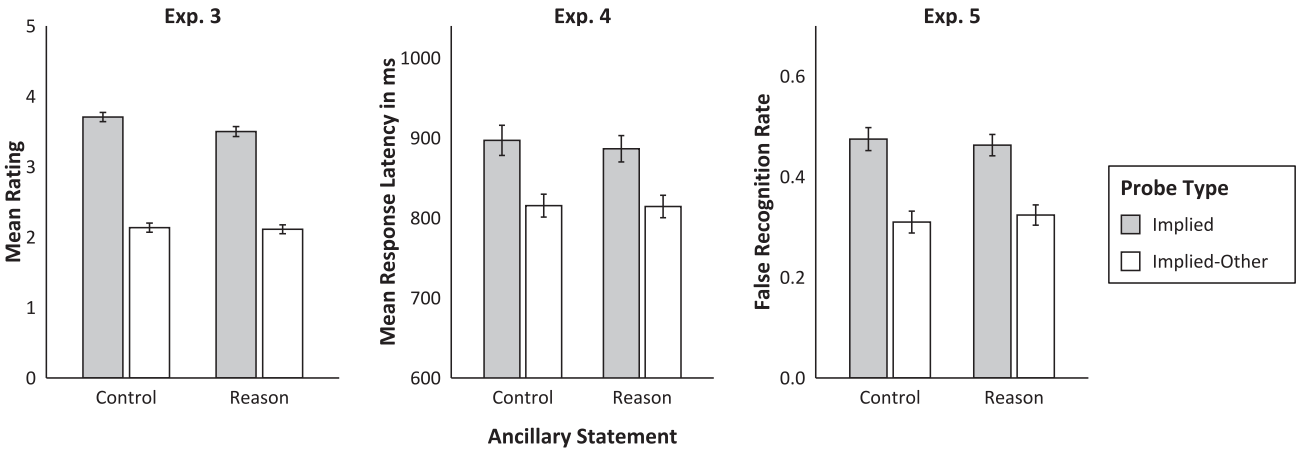


FIGURE 1 Results of Experiments 3, 4, and 5. Mean ratings (Exp. 3), mean response latencies (Exp. 4), and false recognition rates (Exp. 5) are shown for implied and implied-other probes and for control or reason ancillary statements. Error bars show 95% confidence intervals.

15598816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6.0.4 (Campbell & Thompson, 2012), we calculated that we would need a sample of $n = 352$ to obtain 80% power for detecting this effect at $\alpha = .05$ using a repeated-measures ANOVA. We again preregistered a sequential testing procedure with two interim analyses at one- and two-thirds of the total sample. New alpha boundaries for these analyses were calculated with the GroupSeq package for R (Pahl, 2018) using an O'Brien-Fleming-type spending function. We failed to preregister futility bounds, but nonetheless decided to stop for futility at the second interim analysis with $n = 236$ valid data sets and thereby departed from our preregistration. We made this decision mainly so as not to further reduce the already small pool of eligible participants (which provided a test power of 63% for detecting the targeted effect). Following our preregistered exclusion criteria, we excluded nine additional participants, because their average correct response times were slower than two standard deviations over the sample mean, one whose recognition performance was lower than 60% in the probe recognition task, three who rated their own data to be unfit for analysis, and one who withdrew consent for data storage and analysis. Participants were compensated 4.50 GBP for completing the 30 min experiment.

8.1.2 | Materials & procedure

In Experiment 4, we employed the same Probe Recognition Paradigm as in Experiment 2 but using the materials from Experiment 3 (stimulus set 2). Filler items were constructed and probes chosen according to the same logic described for Experiment 2 (see Supporting Information: Table S1.4).

8.2 | Results and discussion

Data preparation was identical to Experiment 2. We conducted a preregistered two-way repeated-measures ANOVA to evaluate the effects of probe type and ancillary statement on the mean response latencies. We observed significant main effects for probe type, $F(1, 235) = 449.71$, $p < .001$, $\eta^2_p = .657$. However, the interaction with ancillary statement did not reach significance, $F(1, 235) = 1.44$, $p = .232$, $\eta^2_p = .006$. We again employed the same additional multiverse analysis as in Experiment 2. The effect of probe type was highly significant across all combinations of cut-off criteria and transformations (all $ps < .001$), with effect size estimates ranging from $\eta^2_p = .592$ to $\eta^2_p = .737$ (see Supporting Information: Table S5.6). The interaction with ancillary statement, however, failed to reach significance across all combinations (ps ranging from .032 to .990) at the preregistered alpha boundary of $\alpha = .012$.

We conclude that participants draw SIIs while reading the statements. Despite the power for detecting a small interaction effect being somewhat reduced due to our decision to stop data collection early, it seems safe to conclude that SIIs were not meaningfully reduced when alternative reasons for the observed behaviors were presented. Therefore, Experiment 4 provides further evidence that SIIs are subject to the correspondence bias. There is, however, an important limitation of the probe recognition paradigm, which applies to interpretation of Experiment 2 and 4. Specifically, one could argue that results are rather a side-effect of mere text-comprehension or associative processing rather than actor-specific person inferences (e.g., reading about soy milk may automatically activate the association vegan without requiring the specific inference that a Michelle who refuses to drink coffee without soy milk is a vegan; see also Orghian et al., 2019). This may also explain why we did not observe a significant discounting effect in Experiment 4: The situational correction of an SII—should it occur—might not necessarily lead to an immediate deactivation of the inferred concept. Thinking that lactose-intolerant Michelle might not be a vegan after all still constitutes thinking about the concept vegan. Therefore, the reasons may have influenced the impressions formed about the actor but without reducing concept activation and thus without being detectable in the probe recognition paradigm. To know whether an impression was formed and included an ideological inference about the actor, we need to demonstrate that ideological inferences are actor-specific. This also allows firmer conclusion with regard to the application of the discounting principle to SIIs. In Experiment 5 we employ another indirect measure addressing this issue.

9 | EXPERIMENT 5

In the fifth experiment, we replicated Experiment 4 using a different indirect measure, namely the False Recognition Paradigm (e.g., Todorov & Uleman, 2002). The aim of this experiment was to show that SII effects truly reflect participants' person impressions rather than mere concept activations and to test whether these impressions are affected by the discounting principle. The main difference to the Probe Recognition Paradigm employed in Experiments 2 and 4 is that the learning and test phase are separated in time and that actors are represented by additional images which later serve as retrieval cues. In the learning phase, participants memorize a series of behavioral descriptions, each combined with the respective actor's face. In the test phase, these faces are presented again, paired with either the ideological category implied by the actor's behavior or an implied-other category implied by the behavior of a different actor. Similar to the probe recognition paradigm, participants indicate via keypress whether the probe word had been part of the actor's statement or not. If participants infer an ideological category and truly represent it as an attribute of the actor, then the actor's face should cause retrieval of this category, thus giving them the false impression that it appeared in the statement. This should manifest in higher rates of false recognitions for implied compared to implied-other categories. Different to the probe recognition paradigm, this effect is unlikely to be driven by current activation of any text-based activated ideological concepts, but by the retrieval of person-specific inferences (see Todorov & Uleman, 2002 for a more detailed discussion).

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

9.1 | Method

9.1.1 | Sample

Applying the same rationale as in Experiment 4, we set our smallest effect size of interest of $f = 0.15$, which would require a sample of $N = 352$ to obtain 80% power for detecting this effect at $\alpha = .05$ using a repeated-measures ANOVA. We preregistered a sequential testing procedure with two interim analyses at one and two-thirds of the total sample. Alpha boundaries and futility boundaries were calculated like in Experiment 3 (but with gammaA and gammaB set to 2). The total sample size increased to $N = 386$. As the effect of interest was neither significant nor exceeding the futility bounds at the interim analyses, we collected the full sample and performed the final analysis with $\alpha = .041$. Following our preregistered exclusion criteria, we excluded seven additional participants because they rated their own data to be unfit for analysis and one who did not give informed consent. We further excluded data of 20 participants due to a programming error. Participants were compensated 3.00 GBP for completing the 20 min experiment.

9.1.2 | Materials

We used the same behavioral descriptions and reasons as in Experiments 3 and 4 (stimulus set 2) but included only 12 filler statements (see Supporting Information: Table S1.5) as is customary for the false recognition paradigm. For the faces, we used AI-generated portrait pictures from the Academic Data set by Generated Photos (https://generated.photos/datasets), selecting pictures such that characteristics like age, gender, and racialized group membership did not seem incongruent with the respective statement.

9.1.3 | Procedure

In Experiment 5, we used the false recognition paradigm (e.g., Todorov & Uleman, 2002). The paradigm has two phases. In the learning phase, participants were presented with 24 target items and 12 filler items in a randomized order (preceded by one practice item) and instructed to memorize the materials. Each target item consisted of a face, an ideology-implying statement and an ancillary statement either containing a reason or not. The ancillary statement (reason vs. control) was manipulated within participants using a Latin square design, such that every participant saw 12 target items in each condition and every item was presented equally often in each condition across all participants. For this experiment we chose a fixed order of presentation with the reason always preceding the ideology-implying statement, because in Experiment 4 the discounting effect had been descriptively larger in this condition (see Supporting Information: S5.8). Each learning trial began with the presentation of a face for 1000 ms, after which both the ideology-implying statement and reason appeared. The duration of presentation was minimally longer than in Experiment 4 (375 ms per word) and ranged from 8.25 to 23.25 s. Intertrial intervals lasted for 750 ms.

In the test phase, participants were told that they would again see the faces from the learning phase—each paired with a single probe word. They were asked to indicate whether the word had appeared in the statement or not by pressing [J] for yes or [F] for no, respectively. After performing two practice trials, they completed 24 target trials and 12 filler trials in an individually randomized order. For half of the target trials the probe was the implied category or an implied-other category, respectively. We created two parallel versions of the stimulus set. In version A, the first half of the 24 target faces were paired with implied labels and the second half with implied-other labels, whereas in version B it was the other way around. Participants were randomly assigned to one of the two versions. The faces from the 12 filler items were presented with category probes that were actually contained in the respective statements. In each trial, the face was presented together with the probe until a response was made. After an intertrial interval of 750 ms, the next test trial began.

9.2 | Results and discussion

We calculated false recognition rates by dividing the number of erroneous responses by the number of trials separately for each experimental condition. We conducted a preregistered three-way mixed-measures ANOVA to evaluate the effects of within-factors probe type, ancillary statement, and stimulus set version as a covariate on the false recognition rates. As expected, we observed a significant main effect for probe type, $F(1, 384) = 221.48$, $p < .001$, $\eta^2_p = .366$. The interaction with ancillary statement was again not significant, $F(1, 384) = 1.70$, $p = .193$, $\eta^2_p = .004$. We observed unexpected effects of the stimulus set version: a main effect, $F(1, 384) = 8.40$, $p = .004$, $\eta^2_p = .021$, a two-way interaction with probe type, $F(1, 384) = 17.99$, $p < .001$, $\eta^2_p = .045$, and a three-way interaction with probe type and ancillary statement, $F(1, 384) = 5.47$, $p = .020$, $\eta^2_p = .014$. Posthoc tests revealed strong inference effects for both stimulus set versions and a small but significant discounting effect for only one of the two versions (for details see Supporting Information: S5.9).

The results from Experiment 5 replicate SIIs in another experimental paradigm, demonstrating that the spontaneously inferred categories are actor-specific. This lends support to the idea that the ideological categories were truly ascribed to the actor and not merely activated momentarily in response to their behavior. Moreover, Experiment 5 provides further evidence that SIIs are subject to the correspondence bias. Evidence for discounting was weak and observable only for one stimulus set version. Overall, providing alternative reasons for behaviors again did not prevent participants from inferring ideological categories.

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10 | GENERAL DISCUSSION

Across five experiments and three different measures, we demonstrated that people deliberately, as well as spontaneously infer ideological categories from ideology-implying behavioral descriptions. In Experiments 1 and 3 we used a direct measure to show that actors performing ideology-implying behaviors were rated as belonging to the implied ideological categories. In Experiments 2, 4 and 5, we used two indirect measures to show that these ideological categories are also inferred spontaneously, that is when people are not forced to make any inferences and are unlikely to have the intention to form an impression. Results of Experiment 5 further support the conclusion of actor-specific ideological inferences. Moreover, across three experiments we found that both deliberate, as well as spontaneous ideological inferences were subject to the correspondence bias: Presenting reasons for the ideology-implying behaviors reduced but not eliminated ideological inferences on a direct measure and had no reliable effect on ideological inferences assessed with indirect measures. Overall, deliberate and spontaneous ideological inferences proved to be very strong and robust.

The robustness of the inferences against additional reasons echoes past research on the correspondence bias, according to which people often draw dispositional inferences from behavior even when it is situationally constrained (Gawronski, 2004). According to classic models of social inference, the correspondence bias might occur because drawing dispositional inferences from behavior is a relatively automatic process whereas the situational correction of a trait inference requires motivated and effortful processing (Gilbert, 1989). Our data fits this account, as we found discounting of ideological inferences only on direct measures, which are more likely to have elicited motivated and effortful processing. While our results suggest that ideological inferences occur without intention and might not be easily controlled, more research is needed to fully understand the extent to which they meet all criteria of automaticity (awareness, intention, efficiency, and control; Bargh, 1994).

Similarly, more research is needed to fully understand the conditions under which discounting of person attributions takes place. Recent research has shown that STIs are not always the dominant inference and can even occur in parallel with spontaneous inferences of mental states and situational constraints (Kruse & Degner, 2021; Todd et al., 2011). For instance, the behavior "Leo cannot get the computer started"; Ham & Vonk, 2003, p. 445) was shown to be as likely to elicit situational inferences (e.g., the computer is broken) as trait inferences (e.g., Leo is inept). This finding challenges the idea of situational correction being a more effortful process per se: If both situational and trait inferences occurred in parallel, attribution would be merely a matter of selecting one of the two. There is no theoretical reason why selecting a trait over a situation should be more effortful than the reverse. In principle, this should apply to ideological inferences as well, which makes it surprising that we did not observe any discounting effects when providing participants with information that provided alternative, situational attributions of behavior. This apparent contradiction might be explained by considering the specific characteristics of behavioral statements used in this research. Whether a behavior elicits multiple inferences, a single trait, or an ideological inference might depend on that behavior's ambiguity. Whereas the stimuli used by Kruse and Degner (2021) and Todd et al. (2011) were selected for their ambiguous implications, our materials were selected to be clearly and unambiguously implying ideological categories. They were thus likely to be perceived as highly diagnostic of those categories. For behaviors with high perceived diagnosticity situational, correction might indeed be more effortful despite the fact that situational inferences can theoretically occur in parallel with trait or ideological inferences. Had we used less diagnostic (i.e., more ambiguous) behaviors, we might have observed smaller inference effects and stronger discounting effects. Although we believe that our materials are realistic and therefore provide a good estimate for the robustness of ideological inferences in everyday life, future research should systematically investigate the effects of perceived diagnosticity on spontaneous person inferences.

If the occurrence and robustness of ideological inferences can be attributed in part to high perceived diagnosticity of the examined behaviors, the question arises as to what renders these behaviors so diagnostic and whether the inferences are justified. Perceived diagnosticity can be formalized as the perceived probability of behavior B being performed by someone who has disposition D divided by the perceived probability of behavior B being performed by someone without disposition D (e.g., Skowronski & Carlston, 1987). To the extent that people can accurately gauge these two probabilities and actual diagnosticity is high, ideological inferences can be considered justified. Findings from the polarization literature suggest that in some contexts actual diagnosticity of political opinions may in fact be high and increasing. In the United States, there is evidence of partisan sorting over the last decades, with conservatives sorting into the Republican Party and liberals sorting into the Democratic Party (Mason & Wronski, 2018; Mason, 2015). This trend is not limited to symbolic ideology (i.e., self-identification as liberal vs. conservative) but extends to operational ideology (i.e., the actual opinions people have regarding specific issues; Park, 2018). Although even in the United States sorting of operational ideology is far from complete (Park, 2018), this nevertheless suggests that actual diagnosticity is increasing, with some opinions regarding strongly politicized issues seemingly becoming highly diagnostic of partisanship (see Lüders et al., 2023). However, it is questionable whether people gauge diagnosticity accurately. In fact, the literature on false polarization suggests otherwise: Differences between political groups are often overestimated (Ahler & Sood, 2018; Druckman et al., 2022; Westfall et al., 2015). People perceive both the political ingroup and outgroup to have more extreme average opinions than they actually do (Ahler, 2014; Levendusky & Malhotra, 2016). This implies that the diagnosticity of these opinions is overestimated. For instance, if a Democrat thought that >1% of Democrats are pro-life but in reality, it is 20%, then they should overestimate the likelihood of a pro-lifer being Republican. This suggests that—at least in the US context—ideological inferences may be to some degree inaccurate.

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

With regard to non-party categories like feminist, racist, or environmentalist the application of the concept of diagnosticity is less straightforward, because these categories have fuzzy boundaries. We propose that a behavior should be considered highly diagnostic of one of these categories, if most behaviors typically associated with it are in alignment. For example, if all behaviors that are typically construed as feminist (e.g., criticizing the gender pay gap, complaining about mansplaining, and so on) are correlated with one another, then these behaviors are diagnostic of the category feminist. If, however, these behaviors are cross-cutting, then one behavior is not predictive of the others such that diagnosticity is low. Echoing the classic finding of low ideological constraint in US citizens' political attitudes (Converse, 2006), Baldassarri and Gelman (2008) report that issue alignment in the general public is rather low and does not seem to be increasing substantially. This suggests that actual diagnosticity for non-party categories may also be rather low. However, they further report that people with greater political interest show stronger issue alignment and a more pronounced increase over time. To the extent that politically interested people have greater public visibility, they may skew perceptions of issue alignment, thus causing overestimation of diagnosticity. Taken together, for both party and non-party categories it is feasible that behaviors' diagnosticity is sometimes overestimated, suggesting that ideological inferences may involve a certain degree of inaccuracy.

Irrespective of their accuracy, SIIs may play an important role in political polarization because many negative outcomes associated with affective and perceived issue polarization should only occur to the extent that people ideologically categorize each other (e.g., as mutual out-partisans). Our research provides strong evidence that this categorization can occur via behavior-based ideological inferences. Given that they appear to occur somewhat automatically, it seems highly likely that such inferences could happen frequently in everyday life. This may be problematic insofar as frequent categorization can lead to frequent ascription of negative attributes such as extreme attitudes (perceived issue polarization) and negative valence (affective polarization). The resulting negative impression may then drive negative interpersonal outcomes, such as co-partisan preferences in online dating or feeling uncomfortable with the idea of out-partisan children-in-law, out-partisan friends, or political discussions with out-partisans (Broockman et al., 2020; Huber & Malhotra, 2017; Iyengar et al., 2012; Pew Research Center, 2017). Thus, to the extent that people arrive at ideological categorizations via spontaneous behavior-based ideological inferences, these inferences should act as a precondition for the negative interpersonal outcomes associated with affective and perceived issue polarization. As both the strength and robustness of the inferences are likely a function of the perceived diagnosticity of the underlying behaviors, future research should investigate if and how perceived diagnosticity of behaviors for ideological categories can be reduced—either in concert with actual diagnosticity or independent thereof.

Although we deem the proposed mechanism by which SIIs may contribute to political polarization highly plausible, further research will need to test it directly. So far, experimental research on the outcomes of different components of political polarization is relatively scant (Broockman et al., 2020). Future research needs to investigate potential further outcomes of affective and perceived issue polarization and whether the outcomes known so far correspond to meaningful real-life behavior (e.g., contact avoidance). Moreover, the role these outcomes may play in the maintenance or amplification of political polarization should be investigated. For instance, contact avoidance might prevent corrective experiences such as realizing that out-partisans have less extreme opinions than presumed, thus contributing to the maintenance of false polarization. To the extent that ideologically categorizing a person leads to experiences of hostility or rejection, this might even push them towards identifying more with the inferred category, which could in turn increase affective polarization. Research investigating how such individual-level processes interact with developments at the societal level would help further gauge the importance of ideological inferences for the process of political polarization.

11 | CONCLUSION

Applying social-cognitive research from the domain of impression formation to the study of political polarization, we presented a comprehensive investigation of (spontaneous) ideological inferences. We outlined a mechanism by which these inferences may affect negative outcomes of political polarization. With this proof of concept, we believe to make an important contribution to the understanding of political polarization. Further research is needed to better comprehend the antecedents, boundary conditions, and consequences of behavior-based ideological inferences and to test potential ways of reducing these inferences. In a world where ideological categories are inferred from even relatively apolitical behaviors, reducing ideological inferences might be a promising avenue for depolarizing interventions. More generally, we encourage viewing political polarization through the lens of person perception and investigating political identity not just as something that is known by perceivers but as something that is actively inferred.

ACKNOWLEDGMENTS

This research was supported by a doctoral scholarship of Studienstiftung des Deutschen Volkes awarded to Carsten Sander. All procedures were approved by the faculty's local ethics committee (protocol number: 2022_006). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Before data collection all experiments and analyses were preregistered with the Open Science Foundation (https://osf.io/f5bv9), where all materials, methods, data, and analysis scripts are publicly available. The preregistration adheres to the disclosure requirements maintained by the Center for Open Science.

15598816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ORCID

Carsten W. Sander http://orcid.org/0000-0003-4032-1338
Juliane Degner http://orcid.org/0000-0002-6634-8099

REFERENCES

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *The Journal of Politics*, *76*(3), 607–620. https://doi.org/10.1017/S0022381614000085

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, *80*(3), 964–981. https://doi.org/10.1086/697253

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, *114*(2), 408–446. https://doi.org/10.1086/590649

Bargh, J. A. (1994). The four horsemen of automaticity: Intention, awareness, efficiency, and control as separate issues. In R. Wyer & T. Srull (Eds.), *Handbook of social cognition: Basic processes* (Vol. 1, 2nd ed., pp. 1–40). Erlbaum.

Bott, A., Brockmann, L., Denneberg, I., Henken, E., Kuper, N., Kruse, F., & Degner, J. (2022). Spontaneous trait inferences from behavior: A systematic meta-analysis. *Personality & Social Psychology Bulletin*. Advance online publication. https://doi.org/10.1177/01461672221100336

Broockman, D. E., Kalla, J. L., & Westwood, S. J. (2020). Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science*, *67*(3), 808–828. https://doi.org/10.1111/ajps.12719

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

T. Carothers, & A. O'Donohue (Eds.). (2019). Introduction, *Democracies divided: The global challenge of political polarization* (pp. 1–13). Brookings Institution Press.

Converse, P. E. (2006). The nature of belief systems in mass publics (1964). *Critical Review*, *18*(1–3), 1–74. https://doi.org/10.1080/08913810608443650

Core Team, R. (2022). R: A language and environment for statistical computing *[Computer software]*. R Foundation for Statistical Computing. https://www.R-project.org/

Corneille, O., & Lush, P. (2023). Sixty years after Orne's American psychologist article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 83–101. https://doi.org/10.1177/10888683221104368

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2022). (Mis)estimating affective polarization. *The Journal of Politics*, *84*(2), 1106–1117. https://doi.org/10.1086/715603

Enders, A. M., & Armaly, M. T. (2019). The differential effects of actual and perceived polarization. *Political Behavior*, *41*(3), 815–839. https://doi.org/10.1007/s11109-018-9476-2

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernbach, P. M., & Van Boven, L. (2022). False polarization: Cognitive mechanisms and potential solutions. *Current Opinion in Psychology*, *43*, 1–6. https://doi.org/10.1016/j.copsyc.2021.06.005

Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science*, *11*(1), 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, *15*(1), 183–217. https://doi.org/10.1080/10463280440000026

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 189–211). Guilford Press.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38. https://doi.org/10.1037/0033-2909.117.1.21

Ham, J., & Vonk, R. (2003). Smart and easy: Co-occurring activation of spontaneous trait inferences and spontaneous situational inferences. *Journal of Experimental Social Psychology*, *39*(5), 434–447. https://doi.org/10.1016/S0022-1031(03)00033-7

Huber, G. A., & Malhotra, N. (2017). Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics*, *79*(1), 269–283. https://doi.org/10.1086/687533

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests* (0.7.2) [Computer software]. https://CRAN.R-project.org/package=rstatix

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128. https://doi.org/10.1037/h0034225

Koudenburg, N., & Kashima, Y. (2022). A polarized discourse: Effects of opinion differentiation and structural differentiation on communication. *Personality and Social Psychology Bulletin*, *48*(7), 1068–1086. https://doi.org/10.1177/01461672211030816

Kruse, F., & Degner, J. (2021). Spontaneous state inferences. *Journal of Personality and Social Psychology*, *121*(4), 774–791. https://doi.org/10.1037/pspa0000232

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D., Pahlke, F., & Wassmer, G. (2021). Group sequential designs: A tutorial [Preprint]. *PsyArXiv*. [Preprint]. https://doi.org/10.31234/osf.io/x4azm

LeBas, A. (2018). Can polarization be positive? Conflict and institutional development in Africa. *American Behavioral Scientist*, *62*(1), 59–74. https://doi.org/10.1177/0002764218756923

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly*, *80*(S1), 378–391. https://doi.org/10.1093/poq/nfv045

Lüders, A., Carpentras, D., & Quayle, M. (2023). Attitude networks as intergroup realities: Using network-modelling to research attitude-identity relationships in polarized political contexts. *British Journal of Social Psychology*. Advance online publication. https://doi.org/10.1111/bjso.12665

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102*(4), 661–684. https://doi.org/10.1037/a0026790

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization: Partisan sorting and polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology*, *39*, 257–277. https://doi.org/10.1111/pops.12485

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

consequences for democratic polities. *American Behavioral Scientist*, *62*(1), 16–42. https://doi.org/10.1177/0002764218759576

Moskowitz, G. B. (2005). *Social cognition: Understanding self and others* (ix, p. 612). Guilford Press.

Orghian, D., Ramos, T., Garcia-Marques, L., & Uleman, J. S. (2019). Activation is not always inference: Word-based priming in spontaneous trait inferences. *Social Cognition*, *37*(2), 145–173. https://doi.org/10.1521/soco.2019.37.2.145

Pahl, R. (2018). GroupSeq: A GUI-based program to compute probabilities regarding group sequential designs. *R Package Version 1.4.2*. https://CRAN.R-project.org/package=GroupSe

Park, B. (2018). How are we apart? Continuity and change in the structure of ideological disagreement in the American public, 1980–2012. *Social Forces*, *96*(4), 1757–1784. https://doi.org/10.1093/sf/sox093

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pew Research Center. (2017). *The partisan divide on political values grows even wider*. https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/

Shimizu, Y. (2017). Why are negative behaviours likely to be immediately invoked traits? The effects of valence and frequency on spontaneous trait inferences. *Asian Journal of Social Psychology*, *20*(3–4), 201–210. https://doi.org/10.1111/ajsp.12183

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*, 689–699. https://doi.org/10.1037/0022-3514.52.4.689

Somer, M., & McCoy, J. (2018). Déjà vu? Polarization and endangered democracies in the 21st century. *American Behavioral Scientist*, *62*(1), 3–15. https://doi.org/10.1177/0002764218760371

Somer, M., McCoy, J. L., & Luke, R. E. (2021). Pernicious polarization, autocratization and opposition strategies. *Democratization*, *28*(5), 929–948. https://doi.org/10.1080/13510347.2020.1865316

Todd, A. R., Molden, D. C., Ham, J., & Vonk, R. (2011). The automatic and co-occurring activation of multiple social inferences. *Journal of Experimental Social Psychology*, *47*(1), 37–49. https://doi.org/10.1016/j.jesp.2010.08.006

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, *83*(5), 1051–1065. https://doi.org/10.1037/0022-3514.83.5.1051

Uleman, J. S., Rim, S., Adil Saribay, S., & Kressel, L. M. (2012). Controversies, questions, and prospects for spontaneous social inferences. *Social and Personality Psychology Compass*, *6*(9), 657–673. https://doi.org/10.1111/j.1751-9004.2012.00452.x

Wassmer, G., & Pahlke, F. (2020). *rpact: Confirmatory adaptive clinical trial design and analysis*. https://www.rpact.org

Wells, B. M., Skowronski, J. J., Crawford, M. T., Scherer, C. R., & Carlston, D. E. (2011). Inference making and linking both require thinking: Spontaneous trait inference and spontaneous trait transference both rely on working memory capacity. *Journal of Experimental Social Psychology*, *47*(6), 1116–1126. https://doi.org/10.1016/j.jesp.2011.05.013

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, *10*(2), 145–158. https://doi.org/10.1177/1745691615569849

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sander, C. W., & Degner, J. (2024). Investigating mechanisms of political polarization: Perceivers spontaneously infer ideological categories from other people's behavior. *Journal of Applied Social Psychology*, *54*, 5–17. https://doi.org/10.1111/jasp.13011

15591816, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jasp.13011 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

CHAPTER 3

From Partisans to Individuals: Lowering an Opinion's Diagnosticity via Counterstereotypes Can Reduce Categorization Based on That Opinion

Sander, C. W., & Degner, J. (2026). From partisans to individuals: Lowering an opinion's diagnosticity via counterstereotypes can reduce categorization based on that opinion. *Political Psychology*, *47*, Article e70130. https://doi.org/10.1111/pops.70130

Check for updates

DOI: 10.1111/pops.70130

ORIGINAL ARTICLE

Political Psychology

From partisans to individuals: Lowering an opinion's diagnosticity via counterstereotypes can reduce categorization based on that opinion

Carsten W. Sander | **Juliane Degner**

Department of Social Psychology, Universität Hamburg, Hamburg, Germany

Correspondence
Carsten W. Sander, Department of Social Psychology, University of Hamburg, Von-Melle-Park 5, Raum 3025a, Hamburg 20146, Germany.
Email: carsten.sander@uni-hamburg.de

Funding information
Studienstiftung des Deutschen Volkes

Abstract

Research has shown that individuals often infer political categories like Republican, Democrat, conservative, or liberal from behavioral cues. In polarized contexts, where political outgroups are evaluated very negatively, these political categorizations may lead to negative outcomes and exacerbate polarization. The current research investigated whether political categorization can be reduced via counterstereotypical exemplars. Across three experiments, we exposed participants to stereotype-confirming vs. stereotype-disconfirming category exemplars to make political opinions seem more or less diagnostic of different political categories. Our findings provide initial evidence that counterstereotypes that decrease the diagnosticity of an opinion can significantly reduce political categorization on the basis of this opinion. This reduction was more pronounced when making a cue appear more common in a category where it is rare, compared with making a cue appear less common in the category where it is frequent. We discuss these findings in light of their potential implications for polarization dynamics. We also consider potential reasons why the diagnosticity of political opinions is perceived as particularly high in polarized contexts and suggest that greater visibility of counterstereotypical individuals may help reduce political categorization and its negative consequences.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2026 The Author(s). *Political Psychology* published by Wiley Periodicals LLC on behalf of International Society of Political Psychology.

INTRODUCTION

Political polarization is increasingly understood through the lens of social identity theory (Iyengar et al., 2012, 2019; Mason & Wronski, 2018; Tajfel & Turner, 1979). High levels of distrust and dislike between supporters of opposing parties are seen not merely as the result of ideological disagreement, but as a consequence of social identification with political groups—such as liberals vs. conservatives, or Democrats vs. Republicans. When people identify with a political group, they may also become motivated to determine whether others belong to the same group or to the opposing side. And indeed, recent research shows that individuals readily infer others' political affiliations based on short descriptions of single behaviors, such as conforming to politically relevant social norms or expressing a political opinion (Sander & Degner, 2024).

In polarized contexts, where political outgroups are evaluated very negatively (Iyengar et al., 2012, 2019), inferring someone's political affiliation can have severe social consequences—not least by enabling the application of category stereotypes (see Myers, 2023; Rothschild et al., 2019). Categorizing someone as a political outgroup member has been shown to reduce trust, empathy, and our willingness to interact with them (Broockman et al., 2023; Carlin & Love, 2013; Simas et al., 2020). In the long run, these consequences may exacerbate polarization by fostering intergroup conflict and impeding constructive dialogue between people with different opinions. Unfortunately, such inferences often occur automatically and are difficult to suppress (Sander & Degner, 2024), and may therefore persist even when individuals are encouraged to avoid them. This raises an important question: Is it possible to prevent these inferences from arising in the first place? The impression formation literature suggests that behavioral cues elicit social category inferences when they are perceived as diagnostic of the relevant category (Skowronski, 2002). From this perspective, reducing the perceived diagnosticity of politically relevant cues may be a promising strategy for dampening political categorization. The present research investigates the feasibility of this approach.

Perceived diagnosticity and political categorization

In many intergroup domains, categorization is driven predominantly by visual cues. For instance, we spontaneously categorize others into racial, gender, or age categories merely by looking at their faces (Colombatto et al., 2021; Mandalaywala & Legaspi, 2023). For political categorization this is different. Unless someone deliberately expresses their identity via visual symbols, we cannot detect their political orientation just by looking at them. In most cases, political categories thus have to be inferred from behavioral cues, and the impression formation literature suggests that we accomplish this by focusing on cues that we perceive as "diagnostic." Diagnosticity describes "the extent to which a cue simultaneously predicts membership in one category and non-membership in the other" (Skowronski, 2002, p. 138). In other words, a cue is diagnostic of a category if the likelihood of a member of category A exhibiting the cue is high and the likelihood of a member of category B exhibiting the cue is low. According to Skowronski and Carlston (1987, p. 692), this idea can be formalized as the following:

$$Diagnosticity = \frac{P(cue\ C|\ category\ A)}{P(cue\ C|\ category\ A) + P(cue\ C|\ category\ B)} \quad (1)$$

Formula (1) shows that diagnosticity depends on the perceived likelihood of people belonging to category A and category B exhibiting cue C. The beliefs about these likelihoods can be viewed as stereotypes—commonly defined as beliefs about attributes and behaviors of members of different categories (Kite et al., 2022). Therefore, reducing a cue's diagnosticity corresponds to changing

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the category stereotypes in such a way that the perceived probabilities of observing the cue in the two categories converge. The idea that categorization is guided by diagnosticity is widely applied in the literature on trait inference (Cone & Ferguson, 2015; Degner, 2025; Skowronski, 2002) and is also implicit in rational models of stereotype change, which describe categorization into social categories as a Bayesian inference process (Gershman & Cikara, 2023; Spicer & Sanborn, 2017). These models predict that people derive probabilistic category inferences from how well an individual's features fit the feature distributions of different groups. Changing group stereotypes in a way that reduces between-group and increases within-group variability regarding a cue should thus lead to reduced categorization on the basis of that cue.

However, not every stereotype is diagnostic. If a cue is perceived as very likely in category A but also moderately likely in category B, then it is very *prototypical* of category A but not very *diagnostic* of it. If, in contrast, a cue is only somewhat likely in category A but completely unlikely in category B, then it is very *diagnostic* of category A despite not being *prototypical* for it (see Chin-Parker & Ross, 2004). To illustrate, between the two stereotypes that Democrats are dishonest and that they want to defund the police, only the latter is likely to be perceived as diagnostic, as only the latter is perceived as *uniquely* tied to Democrats, whereas dishonesty is seen as somewhat typical for politicians in general. Wanting to defund the police may nevertheless be perceived as not prototypical of Democrats, because only a small percentage of Democrats endorse this position. Thus, to decrease political categorization one must identify cues that are diagnostic of political categories and change the stereotypes about these cues in such a way that their likelihoods in both categories become more similar.

Diagnostic cues in polarized societies

One type of cue that is central to our representations of political categories is people's political opinions (Myers, 2023). Political opinions are inherently somewhat diagnostic of political categories, since these categories are at least to some extent formed around opposing opinions. In polarizing societies, however, they often come to be perceived as even more diagnostic. One reason for this is that polarization is characterized by partisan sorting and opinion alignment, which increases the correlation between multiple cues. In the United States, for example, recent decades have seen individuals identifying as conservative vs. liberal sort into the Republican vs. Democratic parties, respectively (Mason, 2015; Mason & Wronski, 2018). This sorting process has made cues of ideological identity more diagnostic of partisanship. The same is true for people's actual opinions, with an increasing share of U.S. Americans holding consistently liberal and conservative views (Pew Research Center, 2014). Although sorting with respect to opinions is far from complete (Park, 2018) and U.S. citizens still seem to be characterized by relatively low ideological constraint (Baldassarri & Gelman, 2008; Converse, 1964/2006), there are nevertheless at least some hot-button issues, that highly align with one another and have become very diagnostic of political categories (Lüders et al., 2023).

A second reason for high perceived diagnosticity is that polarization is often characterized by exaggerated assumptions about the members of political categories (Ahler & Sood, 2018; Fernbach & Van Boven, 2022; Lees & Cikara, 2021). For instance, in polarized contexts, both political ingroups and outgroups are perceived to hold more extreme views than they actually do (Ahler, 2014; Levendusky & Malhotra, 2016). People also underestimate the diversity of opinions within each party (Dias et al., 2025). These misperceptions inflate the perceived diagnosticity of political opinions: if opinions are believed to be both extreme and homogeneous, the perceived overlap between the two parties shrinks, making each opinion seem more predictive of party identity. For example, if all Democrats were thought to be strongly "pro-choice" and all Republicans strongly "pro-life," then abortion stance would appear to perfectly signal party membership.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In short, polarized contexts heighten the perceived diagnosticity of political opinions in two ways: by increasing their *actual* association with party membership and by fostering *exaggerated beliefs* about their extremity and homogeneity. This increased perceived diagnosticity likely leads to more frequent political categorization. In the current research, we thus focus on political opinions in a polarized context—the U.S. context—as our diagnostic cues.

Counterstereotypes

We have argued that reducing a cue's diagnosticity involves shifting category stereotypes such that the perceived likelihood of observing the cue in each category becomes more similar. One well-established method for changing stereotypes is exposure to counterstereotypical exemplars—individuals who belong to a category but behave in ways that contradict its associated stereotypes (Mastro & Tukachinsky, 2011; McIntyre et al., 2016; Ramasubramanian, 2011). If a counterstereotype is assimilated into a group, this should shift the group's stereotype toward the counterstereotypical cue and increase perceived within-group variability. As a result, a novel individual exhibiting this cue should be less likely to be assigned to the group for which this cue is typical (Gershman & Cikara, 2023; Spicer & Sanborn, 2017).

Recent research has demonstrated the effectiveness of counterstereotypes for shifting even entrenched partisan stereotypes. For example, exposure to a Democratic member of Congress who voted against increasing the minimum wage, or a Republican member who voted for mandatory background checks for gun sales, has been shown to shift perceptions of the typical stance of these two categories (Myers & Hvidsten, 2024). However, no research has yet investigated whether through shifting diagnostic stereotypes, counterstereotypes can also intervene at an earlier stage—by preventing people from making political categorizations in the first place. This distinction is important because the negative effects of categorization go above and beyond stereotyping (Dunham, 2018; Essien et al., 2022). The current research therefore investigates whether counterstereotypes designed to reduce the diagnosticity of specific political opinions can lead to reduced categorization on the basis of these opinions.

The effectiveness of counterstereotypes depends on several boundary conditions. Most importantly, stereotypes often resist change because highly atypical exemplars are mentally excluded from the category—a process known as subtyping (Hewstone et al., 1994; Kunda & Oleson, 1995; McIntyre et al., 2016). In other words, if an exemplar deviates too dramatically from what people expect from the respective group (i.e., the group prototype), they may construe the individual as an exception without updating the stereotype. Thus, to maximize stereotype change, counterstereotypical exemplars must strike a balance: They should deviate from the group prototype enough to challenge existing beliefs, but not so much that they are dismissed as unrepresentative outliers (Gershman & Cikara, 2023; Kunda & Oleson, 1997). This can be achieved by presenting exemplars who deviate from their group norm only on a single issue and only to a moderate extent. Additionally, subtyping can be prevented by including metacognitive triggers, such as explicit instructions reasserting the exemplar's group membership (McIntyre et al., 2016).

Category type

Ideological conflict within a society is not necessarily structured along party categories. Especially in multi-party systems, where different parties and their supporters may hold relatively similar views, conflict can instead be organized along broader ideological categories (e.g., conservatives vs. liberals, see Vegetti & Širinić, 2019). It is not clear whether counterstereotypes are equally effective when applied to party and ideological categories. One reason

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to assume that they might not be, is that party categories can easily be defined by a simple behavioral rule (e.g., as people who *vote* for the respective party), whereas ideological categories cannot. This allows for rule-based categorization (see Erickson & Kruschke, 1998; Smith & Sloman, 1994) in the case of party categories, which may shield atypical exemplars from exclusion, thereby increasing the probability of stereotype change. To illustrate, deviating partisans are often linguistically subgrouped as Republicans or Democrats "in name only" ("RINOs" or "DINOs" for short), which suggests that even when they are perceived to have nothing in common with the category, they are still formally included in it. Categorization into ideological categories may—for the lack of a simple behavioral classification rule—be more similarity-based, such that atypical exemplars are more readily excluded. We therefore investigate the effect of counterstereotypes both for party categories and ideological categories.

The current research

In the current research, we investigate whether counterstereotypes designed to reduce the diagnosticity of political opinions can decrease political categorization across three preregistered experimental studies. Specifically, we test the following two hypotheses:

> **H1.** *Exposure to a partisan (Republican or Democrat) expressing a stereotype-disconfirming opinion reduces party categorization based on that opinion, compared to exposure to a stereotype-confirming opinion.*

> **H2.** *Exposure to an ideological category member (Conservative or Liberal) expressing a stereotype-disconfirming opinion reduces ideological categorization based on that opinion, compared to exposure to a stereotype-confirming opinion.*

Hypothesis 1 is tested in Experiments 1 through 3, while Hypothesis 2 is tested in Experiments 2 and 3. Additionally, Experiments 2 and 3 explore the individual contributions of the two components that constitute diagnosticity (i.e., the two underlying probabilities). Across all three experiments, we employ an impression formation paradigm (Kunda & Oleson, 1997), in which participants are presented with information about a category exemplar, that exhibits either a stereotype-confirming or stereotype-disconfirming cue. Subsequently, they are asked about their impression of the individual and about their impression of the category as a whole (under the pretense of wanting to control for general group impressions). Change in categorization is measured by comparing category judgments before and after presentation of the counterstereotypical exemplar (Exp. 1) or between participants that were presented a stereotypical vs. counterstereotypical exemplar (Exp. 2 and 3).

In the context of the current research, the stereotypes relate to political opinions that are diagnostic of a political category (specifically, opinions regarding gun control, environmental protection, and affirmative action). We present these opinions in a transcript of a purportedly real radio show in which two ordinary people discuss politics. In the stereotype-confirming (high-diagnosticity) condition, participants are presented with a category exemplar that expresses an opinion typical of their category (e.g., a Republican opposing gun control measures), whereas in the stereotype-disconfirming (low-diagnosticity) condition, they are presented with a counterstereotype (e.g., a Republican supporting a ban on assault weapons). We then assess political categorization on the basis of the political opinion.

Participants for all three experiments were US Americans recruited via the online platform Prolific (www.prolific.co) and all experiments were run using Qualtrics (www.qualtrics.com). Analyses were performed using R (R Core Team, 2023) and the rstatix package (Kassambara, 2023). We report sample size determination, data exclusions, manipulations,

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and all measures in the description of each experiment and in the Supporting Information. A pilot study in which we failed to find the expected effect—most likely due to methodological issues—is reported in Supporting Information S1. All procedures were approved by the local ethics committee (protocol number: VA_2024_013).

EXPERIMENT 1

In Experiment 1, we tested whether exposure to a member of a party category (Republican or Democrat) expressing a stereotype-disconfirming opinion reduces party categorization based on that opinion, compared with exposure to a stereotype-confirming opinion (H1).

Methods

Sample

We set our smallest effect size of interest for a two-way interaction in a mixed ANOVA to a small effect of $f = .15$. Using G*Power, we calculated a required $N = 582$ valid data sets to attain a test-power of $1 - \beta = .95$ at a significance level of $\alpha = .05$ (Faul et al., 2007). To avoid unnecessary spending in case of a larger than expected effect size, we preregistered a sequential testing procedure (Lakens et al., 2021; for details, see preregistration). Because the effect of interest was significant in the first interim analysis (with $\alpha = .006$), we stopped data collection early at $n = 216$ valid data sets (see Table 1 for sample demographics). Following our preregistered exclusion criteria, we excluded data of 24 additional participants who inferred the wrong category in three or more pre-measurement categorization items (indicating inattention, random

TABLE 1 Demographic data for all samples.

Variable	Exp. 1	Exp. 2	Exp. 3
Gender			
Female	124	180	323
Male	90	175	302
Non-binary	2	4	7
Undisclosed	0	1	0
Age			
Mdn	39	46	47
Range	18–86	19–79	18–83
Partisanship			
Democrat	129	150	254
Republican	38	118	216
Independent	37	82	143
Other	12	7	15
Undisclosed	0	3	4
Political ideology			
M	–	55.6	56.1
SD	–	31.6	30.8

Note: Political ideology was measured using a sliding scale from 0 (very conservative) to 100 (very liberal).

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

responding, a misinterpretation of the items, or a lack of political knowledge). Participants were compensated 1.05 GBP (≈1.31 USD) for completing the 7-min experiment.

Procedure and materials

As part of our cover story, we invited participants to an alleged market research study evaluating a new radio show format, in which each episode features two average citizens discussing their political opinions. The format was described as providing a more realistic impression of political differences compared with the extreme voices usually amplified by social media. In the transcript, one of the guests (the target guest) who identified either as a Republican or Democrat (target partisanship; varied between-subjects) reported either a stereotype-confirming or stereotype-disconfirming opinion (typicality; varied between-subjects) about the target issue.

Before reading the transcript, participants completed a measure of political categorization for six political opinions pertaining to six polarized political issues (pre-measurement), which were shown to be perceived as highly diagnostic (Lüders et al., 2023). Because the original opinions were mostly phrased as typical Democrat opinions, we constructed opposite versions of each opinion (for all opinions, see Supporting Information S2). Across all participants, we manipulated the diagnosticity for opinions on three of these issues (gun control, environmental protection, and affirmative action), whereas the remaining three served as fillers (income equality, immigration, and abortion). For each participant, diagnosticity for only one opinion was manipulated via the transcript (target opinion; varied between-subjects). In the categorization measure, both this target opinion, as well as the two remaining non-filler opinions (control opinions) were phrased as atypical for the target guest's category. Thus, if the target guest was introduced as a Republican [Democrat], the target and control opinions were phrased as the typical opinion of a Democrat [Republican]. The filler opinions, on the other hand, were phrased as typical of the target guest's category. Participants were presented with a sliding scale from 0 ("Definitely a Republican") to 100 ("Definitely a Democrat") for each opinion and given the following instruction: "Please use the slider to indicate how likely you think a person who made each statement is to be a Republican or a Democrat."

Participants then proceeded with reading the transcript in two rounds. In the first round, the radio show's guests answered three quick introductory questions, during which one identified as Democrat and the other as Republican. After this round, participants completed two filler items about how likable they found the guests. In the second round, the guests responded to two political opinion statements. The first opinion statement was always about income equality and both guests responded in stereotype-confirming ways. This served to establish the guests as prototypical category members to prevent subtyping. The topic of the second statement varied between participants (gun control vs. affirmative action vs. environmental protection). While one guest always responded in a stereotype-confirming way, the other (target) guest responded either in a stereotype-confirming or in a stereotype-disconfirming way while reaffirming their partisanship. In the stereotype-disconfirming condition, the target guest additionally expressed the belief that an increasing number of fellow ingroup members shared their view (e.g., "I think there's actually an increasing number of Republicans who see it this way"). This served as an additional meta-cognitive trigger for inclusion of the target into the category, which should further lower the probability of subtyping (McIntyre et al., 2016). For an example of the typicality manipulation, see Figure 1. A complete example transcript can be found in Supporting Information S3.

After reading the transcript, participants again completed the categorization measure (post-measurement), followed by a stereotype measure that served as a manipulation check. For the stereotype measure, participants were presented with a sliding scale from 0 ("strongly disagree")

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



FIGURE 1 Example of the typicality manipulation. The figure shows the last three responses of an example transcript in which the Republican target guest (black boxes) expresses either a stereotype-disconfirming (upper panel) or stereotype-confirming (lower panel) opinion on the target issue (affirmative action). The full transcript is available in Supporting Information S3.

to 100 ("strongly agree") for the same six opinions as the categorization pre-measurement and given the following instruction: "Please indicate how Republicans [Democrats], on average, would respond to the statements below". We chose the stereotype measure as a manipulation check due to its close relationship with the manipulated variable and its established use as an outcome in prior research. Next, participants provided demographic data (age, gender, partisanship), judged whether they considered their data fit for analyses ("yes," "no"), and were asked two suspicion check questions about whether anything about the radio show transcript stood out to them and whether they suspected that they were deceived about any aspect of the study. Ultimately, they were debriefed about the purpose of the study and asked for their

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

informed consent for data storage and analysis. The whole experiment lasted approximately 7 min.

Results

Data preparation

Stereotype scores were recoded such that a value of 100 reflects a stereotypical response (e.g., indicating that a Republican would "strongly disagree" with a pro-gun control opinion) and a value of 0 reflects a counterstereotypical response (e.g., indicating that a Democrat would "strongly agree" with a pro-gun rights opinion). Categorization scores were also recoded, such that a value above 50 reflects a categorization that is in line with the stereotypes (e.g., categorizing someone with a pro-gun control opinion as a Democrat) and a value below 50 reflects an "incorrect" categorization (e.g., categorizing someone with a pro-gun control opinion as a Republican).

Manipulation check

We conducted a one-sided independent samples *t*-test comparing stereotyping in the stereotype-confirming and stereotype-disconfirming conditions. Stereotyping was significantly weaker when the stereotypes were disconfirmed ($M = 66.0$, $SD = 23.4$) as opposed to when they were confirmed ($M = 76.3$, $SD = 25.8$), $t(212.05) = -3.08$, $p = .001$, $d = -.42$, 90% CI [−.67, −.18], indicating that our manipulation successfully changed the party stereotypes about the target opinion.

Party categorization

Our preregistered analysis was a two-way mixed ANOVA with political categorization as the dependent variable, time as a within-subjects factor, and typicality as a between-subjects factor. We obtained significant effects of typicality, $F(1, 214) = 11.77$, $p < .001$, $\eta_p^2 = .052$, 90% CI [.014, .107], and time, $F(1, 214) = 5.42$, $p = .021$, $\eta_p^2 = .025$, 90% CI [.002, .068]. As expected, these effects were qualified by a significant interaction, $F(1, 214) = 28.50$, $p < .001$, $\eta_p^2 = .118$, 90% CI [.058, .186] (see Figure 2). Simple effects analyses revealed that the effect of time was significant and negative in the stereotype-disconfirming condition, $t(107) = -5.06$, $p < .001$, $d_z = -.49$, 95% CI [−.68, −.32], indicating that the counterstereotypical exemplars reduced political categorization compared with the baseline level. In the stereotype-confirming condition, the effect was significant and positive, $t(107) = 2.31$, $p = .023$, $d_z = .22$, 95% CI [.030, .43], indicating that stereotypical exemplars increased political categorization compared with the baseline level. The effect size in the stereotype-disconfirming condition was significantly larger than in the stereotype-confirming condition, $t(209.32) = 2.33$, $p = .021$, $d = .32$, 95% CI [.050, .60], suggesting that the interaction was more strongly driven by the decrease in political categorization due to the counterstereotypical exemplars than by the increase due to the stereotypical ones.

To test whether the effect was specific to the target opinion, we analyzed the effect of typicality and time on the control opinions. That is, we used the average of the two control opinions, for which diagnosticity had not been manipulated for the respective participant, as the dependent variable of a two-way mixed ANOVA with time as a within-subjects factor and typicality as a between-subjects factor. We obtained a small significant effect of typicality, $F(1, 214) = 4.40$, $p = .037$, $\eta_p^2 = .020$, 90% CI [.001, .061]. The effect of time, $F(1, 214) = .96$, $p = .329$, $\eta_p^2 = .004$, 90% CI [.000, .031], and the interaction, $F(1, 214) = 1.42$, $p = .235$, $\eta_p^2 = .007$, 90% CI

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

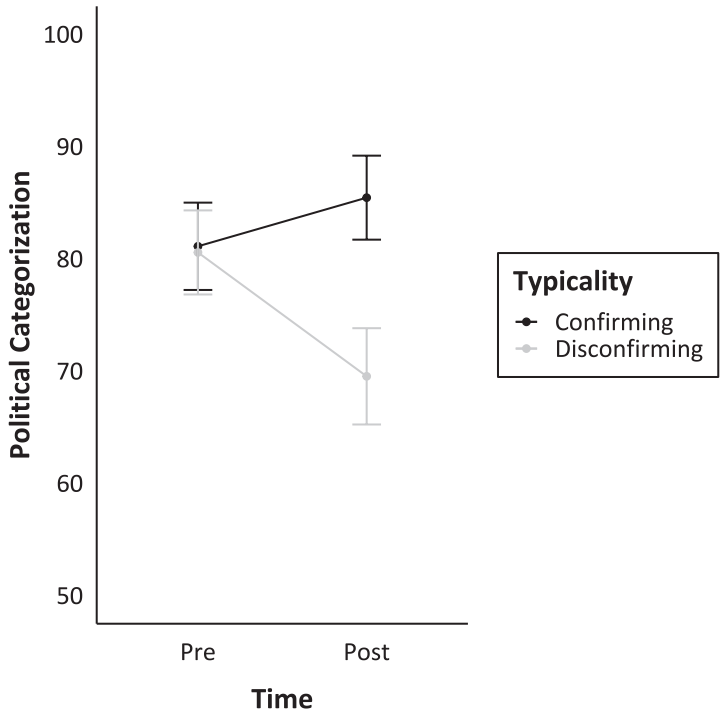


FIGURE 2 Mean categorization scores for Experiment 1. A categorization score of 100 indicates that the person with the target opinion is perceived as definitely belonging to the category associated with that opinion, while a score of 50 indicates they are seen as equally likely to belong to either category. Error bars represent 95% confidence intervals.

[.000, .036], were not significant. This suggests that the effect of our manipulation neither generalized to other opinions nor reflected a mere change in response style.

Discussion

Experiment 1 provided the first evidence for our hypothesis that reducing the diagnosticity of political opinions can reduce political categorization of a hypothetical person holding that opinion. Specifically, presenting participants with a stereotype-disconfirming member of a party category reduced categorization, whereas presenting them with a stereotype-confirming exemplar slightly increased it compared with baseline. One potential limitation of this experiment is that the pre-post design made it fairly easy for participants to correctly guess the research hypothesis. Such demand characteristics could have induced hypothesis-compliant responses. We addressed this limitation in Experiment 2. Another potential limitation is that Republicans were underrepresented in our sample. Although we do not expect that the basic cognitive processes involved in the effects vary based on partisanship, it is possible that our materials would be less effective for Republicans due to varying levels of prior stereotypes or differences in how plausible our manipulation was perceived (for an instance of such diverging effects see Myers & Hvidsten, 2024). We thus drew the samples of Experiments 2 and 3 to be representative of the US population in terms of political affiliation.

EXPERIMENT 2

The main objective of Experiment 2 was to investigate whether the findings from Experiment 1 replicate for ideological categories. We therefore added category type (party category vs. ideological category) as a between-subjects factor, expecting significant effects of typicality also in

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the ideological category condition (H2). Moreover, we addressed the limitations of Experiment 1 by reducing demand characteristics through the use of a between-subjects design (i.e., by removing the pre-measurement) and the use of a red herring cover story (i.e., pretending that we were interested in the impressions participants form of the show's guests and framing the categorization measure as a mere control variable).

An additional objective of Experiment 2 was to explore whether both diagnosticity components affect political categorization. As a reminder, diagnosticity consists of two components: First, the probability of the cue being exhibited by members of the category for which it is common, which we refer to as the *stereotype probability* ($P(cue\ C|\ category\ A)$) and second, the probability of the cue being exhibited by members of the category for which it is uncommon, which we refer to as the *counterstereotype probability* ($P(cue\ C|\ category\ B)$). Perceived diagnosticity can be decreased by increasing the counterstereotype probability or by decreasing the stereotype probability. In the context of the current research, presenting a counterstereotypical exemplar should do both—but for different opinions. For example, a pro-gun control Republican should not only increase the counterstereotype probability for pro-gun control opinions but also decrease the stereotype probability for pro-gun rights opinions—due to the mutual exclusiveness of these opinions. Experiment 1 investigated only the effect of changing the counterstereotype probability, because the categorization measure included only opinions atypical for the target guest's category. In Experiments 2 and 3 we additionally presented the opposite opinions to investigate the impact of both diagnosticity components.

Methods

Sample

Because the current experiment used a between design, we based our sample size calculation on the observed difference between the stereotype-confirming and stereotype-disconfirming conditions in the post-measurements of Experiment 1. We obtained a Cohen's d of −.75 for this comparison but nevertheless set our smallest effect size of interest to a medium effect of −.50 to account for the possibility of smaller effect sizes due to reduced demand effects. To attain a test power of $1 - \beta = .95$ at a significance level of $\alpha = .05$ in two one-sided independent samples t-tests, we required $N = 352$ valid data sets (G*Power; Faul et al., 2007). To ensure equal numbers of participants for each cell of the design, we collected $N = 360$ valid data sets. Following our preregistered exclusion criteria, we excluded data of one additional participant who rated their data as unfit for analysis and 28 (13 in the stereotype-confirming and 15 in the stereotype-disconfirming condition) who inferred the wrong category on at least five of the non-target categorization items.[1] Participants were compensated 1.20 GBP (≈1.52 USD) for completing the 8-min experiment. The sample was drawn to be representative of the US population in terms of sex, age, ethnicity, and political affiliation (see Table 1 for sample demographics).

[1]Because these exclusions were based on measures collected after the manipulation, there is a risk that they could bias the randomization. To assess this, we examined the distribution of exclusions across experimental conditions—jointly for experiments 2 and 3, which used the same methodology. Across both experiments, 52.5% of excluded participants were in the stereotype-confirming condition and 47.5% in the stereotype-disconfirming condition. An exact binomial test showed no significant deviation from a 50:50 distribution, $p = .798$, 95% CI [.346, .607]. This pattern is inconsistent with the idea that the manipulation itself caused the "wrong" categorizations. More plausibly, these resulted inattention, random responding, or a misinterpretation of the items. Supporting this interpretation, the exclusion rates in Experiments 2 (8%) and 3 (5%) were similar to that in Experiment 1 (10%), where the criterion was applied to *pre-manipulation* measures only.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Procedure and materials

The procedure was similar to Experiment 1 except for the following changes: In the beginning of the experiment, we told participants that they would report their impressions of the show's guests after reading the transcript. To increase the plausibility of the cover story, we removed the first categorization measure (pre-measurement) and repeated the likeability measure after the second round of the transcript. Moreover, we added category type (party category vs. ideological category) as a between-subjects factor. Because in the US there are high levels of alignment between partisanship and symbolic ideology (self-categorization as liberal or conservative; Mason, 2015; Mason & Wronski, 2018), we used the same transcripts for both category type conditions and merely swapped the category labels. Accordingly, the question about the guests' category membership varied between conditions: In the party category condition, it was about their political party preference, whereas in the ideological category condition, it was about their ideological self-categorization as conservative vs. liberal.

We altered the cover story of the categorization measure, telling participants that we assess the general typicality of the opinions for the categories because this may affect their impressions of the guests. We also added more scale anchors to the categorization measure to discourage binary interpretations of the scale ("definitely a Republican [Democrat]," "likely a Republican [Democrat]," "equally likely," "likely a Democrat [Republican]," "definitely a Democrat [Republican]") and added a second set of categorization items using the opposite versions of the opinions (stereotype probability condition) on a separate page. The order of the two sets of categorization items was randomized between participants. Next, we added an exploratory measure of perceived prototypicality of both guests, asking participants, "One of the show's participants identified as a Republican [Democrat]. How typical was this participant of Republicans [Democrats] in general?" Participants responded on a sliding scale from 0 (not at all typical) to 100 (very typical). Results for this measure are reported in Supporting Information S4. Lastly, we included a question of political ideology ("Where would you place yourself on the following political spectrum?") with a sliding scale from 0 (very conservative) to 100 (very liberal).

Results

Exclusion of outliers

Inspection of the data revealed several categorization responses that were unlikely to be valid. For 7% of participants, the categorization was opposite to the common group stereotypes (e.g., categorizing someone with a pro-gun control opinion as Republican). Theoretically, any values below the scale midpoint (50) are unlikely to reflect genuine categorizations, because they would indicate that members of the category that is not stereotypically associated with an opinion are more likely to express it than members of the category that is stereotypically associated with it. Even for participants in the stereotype-disconfirming condition, such responses would be implausible, as a singular exposure to a counterstereotype is not expected to reverse a cue's diagnosticity. We therefore decided to exclude these outlying values. While our reasoning would justify excluding values below 50, we adopted a more conservative cutoff of 40 to allow for some inaccuracy in slider use.[2]

[2]To check for potential bias—just as we did for participant exclusions—we analyzed the distribution of outliers across conditions, pooling data from Experiments 2 and 3. Of these outliers, 51.4% came from the stereotype-confirming condition and 48.6% from the stereotype-disconfirming condition. Again, this distribution did not significantly differ from chance ($p = .908$, 95% CI [.369, .606]), supporting the conclusion that these "wrong" categorizations likely reflected factors unrelated to the manipulation, such as inattention, random responding, or a misinterpretation of the items.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A similar problem occurred for the manipulation check measure on which some participants indicated that category members would on average strongly disagree with opinions that are stereotypical for their category. However, for this measure, there is less of a clear theoretical cutoff because even an opinion that is very stereotypical or diagnostic for a category must not be shared by a majority of its members (as long as it is shared by considerably less members of the opposing category). To nevertheless investigate the effect of these outliers, we conducted a multiverse analysis using several different cutoff criteria (see Supporting Information S5). Because we had not pre-registered any data exclusions, we report results of all main analyses with and without applying this exclusion criterion.

Manipulation check

We conducted a two-way between-subjects ANOVA with the stereotype measure as the dependent variable and typicality and category type as between-subjects factors. We applied a multiverse approach with multiple fixed outlier cutoffs excluding data of participants with values below 1, 10, 20, 30, or 40 (on a scale ranging from 0 to 100). The effect of typicality was marginally significant without outlier exclusion ($p = .069$, $\eta_p^2 = .009$) and significant with all different outlier thresholds (all ps $\leq .016$, η_p^2s $\geq .020$, see Supporting Information S5), indicating reduced stereotypes in the stereotype-disconfirming compared with the stereotype-confirming condition. The effect of category type and the two-way interaction were not significant in any of the tests. These results indicate that our manipulation was at least moderately effective in altering category stereotypes about the target opinions, particularly after excluding potentially invalid responses.

Party categorization

Counterstereotype probability

For all of our analyses, we conducted one-sided independent samples t-tests comparing political categorization in the stereotype-confirming and stereotype-disconfirming conditions. The tests for the counterstereotype probability condition were preregistered, whereas the tests in the stereotype probability condition were exploratory. In the counterstereotype probability condition, categorization by party was not significantly weaker when the stereotypes were disconfirmed ($M = 77.6$, $SD = 19.6$) as opposed to when they were confirmed ($M = 81.4$, $SD = 20.9$), $t(177.22) = -1.27$, $p = .104$, $d = -.19$, 90% CI [−.45, .070]. After excluding 9 participants with outlying values, categorization by party was significantly weaker when the stereotypes were disconfirmed ($M = 80.2$, $SD = 15.6$) as opposed to when they were confirmed ($M = 85.3$, $SD = 13.6$), $t(166.51) = -2.27$, $p = .012$, $d = -.35$, 90% CI [−.60, −.090].

Stereotype probability

In the stereotype probability condition, categorization by party was not significantly weaker when the stereotypes were disconfirmed ($M = 80.2$, $SD = 22.2$) as opposed to when they were confirmed ($M = 80.4$, $SD = 24.5$), $t(176.35) = -.064$, $p = .475$, $d = -.010$, 90% CI [−.26, .25]. After excluding 15 participants with outlying values, categorization by party was still not significantly weaker when the stereotypes were disconfirmed ($M = 84.7$, $SD = 14.8$) as opposed to when they were confirmed ($M = 87.4$, $SD = 13.0$), $t(161.53) = -1.23$, $p = .110$, $d = -.19$, 90% CI [−.44, .070].

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ideological categorization

Counterstereotype probability

In the counterstereotype probability condition, ideological categorization was not significantly weaker when the stereotypes were disconfirmed ($M = 79.0$, $SD = 17.4$) as opposed to when they were confirmed ($M = 81.7$, $SD = 20.4$), $t(173.75) = -.96$, $p = .168$, $d = -.14$, 90% CI [−.42, .10]. After excluding 8 participants with outlying values, ideological categorization was significantly weaker when the stereotypes were disconfirmed ($M = 80.1$, $SD = 16.0$) as opposed to when they were confirmed ($M = 86.0$, $SD = 12.9$), $t(165.59) = -2.66$, $p = .004$, $d = -.41$, 90% CI [−.68, −.17].

Stereotype probability

In the stereotype probability condition, ideological categorization was not significantly weaker when the stereotypes were disconfirmed ($M = 76.8$, $SD = 20.9$) as opposed to when they were confirmed ($M = 81.2$, $SD = 19.4$), $t(177.09) = -1.46$, $p = .074$, $d = -.22$, 90% CI [−.48, .020]. After excluding 9 participants with outlying values, ideological categorization was still not significantly weaker when the stereotypes were disconfirmed ($M = 80.5$, $SD = 14.6$) as opposed to when they were confirmed ($M = 84.0$, $SD = 14.6$), $t(168.98) = -1.57$, $p = .059$, $d = -.24$, 90% CI [−.48, .005]. Figure 3 illustrates the mean categorization scores for each condition.

Discussion

The goal of Experiment 2 was to test whether the effects of counterstereotypes on political categorization could be replicated in a design with reduced demand characteristics, and whether they extended to ideological categories. However, the preregistered analyses revealed no significant effect of typicality on political categorization. Inspection of outliers suggested that this may be due to a small number of likely invalid responses that we

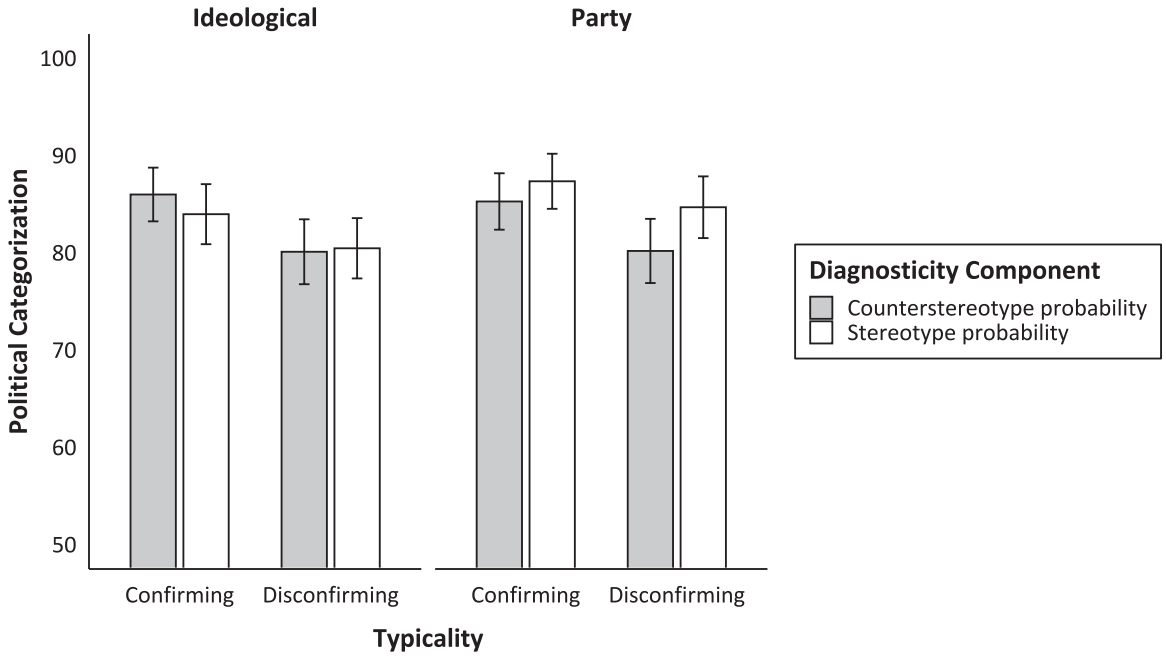


FIGURE 3 Mean categorization scores for Experiment 2. A categorization score of 100 indicates that the person with the target opinion is perceived as definitely belonging to the category associated with that opinion, while a score of 50 indicates they are seen as equally likely to belong to either category. The figure presents results after applying the outlier cutoff (i.e., after excluding values below 40). Error bars represent 95% confidence intervals.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

presume were due to inattention, random responding, or a misinterpretation of the items. When we excluded these responses, our hypotheses were confirmed (for analyses showing that applying the same exclusion criteria to the data of Experiment 1 also led to increased effect sizes, see Supporting Information S6). With this caveat in mind, the results from Experiment 2 provide further, albeit tentative, support that exposure to a category member expressing a stereotype-disconfirming opinion reduces categorization based on that opinion, compared with exposure to a stereotype-confirming opinion. Given that these effects persisted when demand characteristics were reduced by removing the pre-measurement and adding a red herring to the cover story, it is unlikely that they were merely driven by demand effects. Importantly, the results of Experiment 2 suggest that counterstereotypes are effective in reducing categorization both for ideological categories like Liberal and Conservative and for party categories like Democrat and Republican. Provided that these results can be replicated, this speaks to the generalizability of counterstereotypes as an intervention to reduce categorization.

In this experiment, we also investigated categorization on the basis of cues that are typical for the target person's category (stereotype probability condition). Although descriptively the effect went in the expected direction, it was not significant—even when excluding statistical outliers. In other words, presenting participants with a pro-gun control Republican reduced categorization on the basis of pro-gun control opinions, but presenting a pro-gun rights Democrat did not (or only to a very small degree). As we will demonstrate later, this is not surprising as for cues with a high baseline diagnosticity, reducing the stereotype probability is expected to have a smaller effect than increasing the counterstereotype probability by the same amount.

These findings should be regarded as preliminary, as they are based on analyses that deviated from our preregistration and therefore require replication.

EXPERIMENT 3

We planned Experiment 3 as an exact replication of Experiment 2 with the only exception that we preregistered the outlier exclusions of values below 40 on the categorization measures.

Methods

Sample

The sample size was determined using a simulation, based on the effect size of typicality across both category type conditions after exclusion of outliers in Experiment 2 ($d = -.38$). To attain a test-power of $1 - \beta = .95$ at a significance level of $\alpha = .05$ in two one-sided independent samples t-tests, we required valid data from 150 participants per group, that is $N = 600$ participants in total. Because we expected that around 5% of data points in each analysis would be statistical outliers, we collected a sample of $N = 632$ valid data sets. Participants were compensated 1.20 GBP (≈1.56 USD) for completing the 8-min experiment. Following our preregistered exclusion criteria, we excluded 2 additional participants who withdrew their consent to data analysis after full debriefing, 3 participants who rated their data as unfit for analysis, and 33 (19 in the stereotype-confirming and 14 in the stereotype-confirming condition) who inferred the wrong category on at least five of the non-target categorization items. The sample was drawn to be representative of the US population in terms of sex, age, and political affiliation (see Table 1 for sample demographics).

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Procedure and materials

The procedure and materials were identical to Experiment 2.

Results

Exclusion of outliers

For the analyses of the categorization scores, we excluded responses below a threshold of 40 (as preregistered). Analyses without outlier exclusions yielded largely identical results, which are reported in Supporting Information S7. For the stereotype measure, we again conducted a multiverse analysis using several different cutoff criteria.

Manipulation check

To test whether the manipulation was successful, we conducted a multiverse analysis with multiple fixed outlier cutoffs, excluding participants with values below 1, 10, 20, 30, or 40 (on a scale ranging from 0 to 100). We conducted two-way between-subjects ANOVAs with the stereotype measure as the dependent variable and typicality and category type as between-subjects factors for each of the cutoff criteria. The effect of typicality was significant with no outlier exclusion and across all different outlier thresholds ($ps < .001$, $\eta_p^2 s \geq .027$, for the full results of the multiverse analysis see Supporting Information S5). The effect of category type and the two-way interaction were not significant in any of the tests. These results provide clear support for the effectiveness of our manipulation.

Party categorization

Counterstereotype probability

For all of our analyses, we again conducted one-sided independent samples *t*-tests comparing political categorization in the stereotype-confirming and stereotype-disconfirming conditions. The tests for the counterstereotype probability condition were preregistered, whereas the tests in the stereotype probability condition were exploratory. In the counterstereotype probability condition, categorization by party was significantly weaker when the stereotypes were disconfirmed ($M = 78.6$, $SD = 15.5$) as opposed to when they were confirmed ($M = 87.0$, $SD = 12.8$), $t(287.80) = -5.17$, $p < .001$, $d = -.60$, 90% CI [−.80, −.40].

Stereotype probability

In the stereotype probability condition, categorization by party was also significantly weaker when the stereotypes were disconfirmed ($M = 81.3$, $SD = 15.3$) as opposed to when they were confirmed ($M = 85.1$, $SD = 13.6$), $t(289.55) = -2.24$, $p = .013$, $d = -.26$, 90% CI [−.46, −.070].

Ideological categorization

Counterstereotype probability

In the counterstereotype probability condition, ideological categorization was significantly weaker when the stereotypes were disconfirmed ($M = 75.8$, $SD = 14.8$) as opposed to when they were confirmed ($M = 86.5$, $SD = 13.8$), $t(277.93) = -6.26$, $p < .001$, $d = -.74$, 90% CI [−.96, −.52].

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Stereotype probability

In the stereotype probability condition, ideological categorization was also significantly weaker when the stereotypes were disconfirmed ($M = 79.0$, $SD = 16.4$) as opposed to when they were confirmed ($M = 86.0$, $SD = 13.3$), $t(267.54) = -3.99$, $p < .001$, $d = -.47$, 90% CI [−.68, −.29]. Figure 4 illustrates the mean categorization scores for each condition.

Discussion

Experiment 3 supports the conclusions that (a) exposure to category members expressing a stereotype-disconfirming opinion reduces categorization of a hypothetical individual holding that opinion, compared with exposure to a stereotype-confirming opinion, (b) this is true also for ideological categories that lack a clear behavioral categorization rule, and (c) these effects are unlikely to be driven by demand effects. The representative samples used in Experiments 2 and 3 also address the underrepresentation of Republicans in Experiment 1 and additional analyses indicate that partisanship did not moderate the observed effects (see Supporting Information S8). In contrast to Experiment 2, we did find significant effects in the stereotype probability condition. However, these effects were descriptively smaller than in the counterstereotype probability condition. In the following section, we report exploratory analyses on the combined data from Experiments 2 and 3 that test whether the difference between these two conditions was statistically significant.

EXPLORATORY ANALYSES OF EXPERIMENTS 2 AND 3

For exploratory analyses across the data from Experiments 2 and 3, we combined the data of both experiments and again employed the outlier cutoff, excluding values below 40.

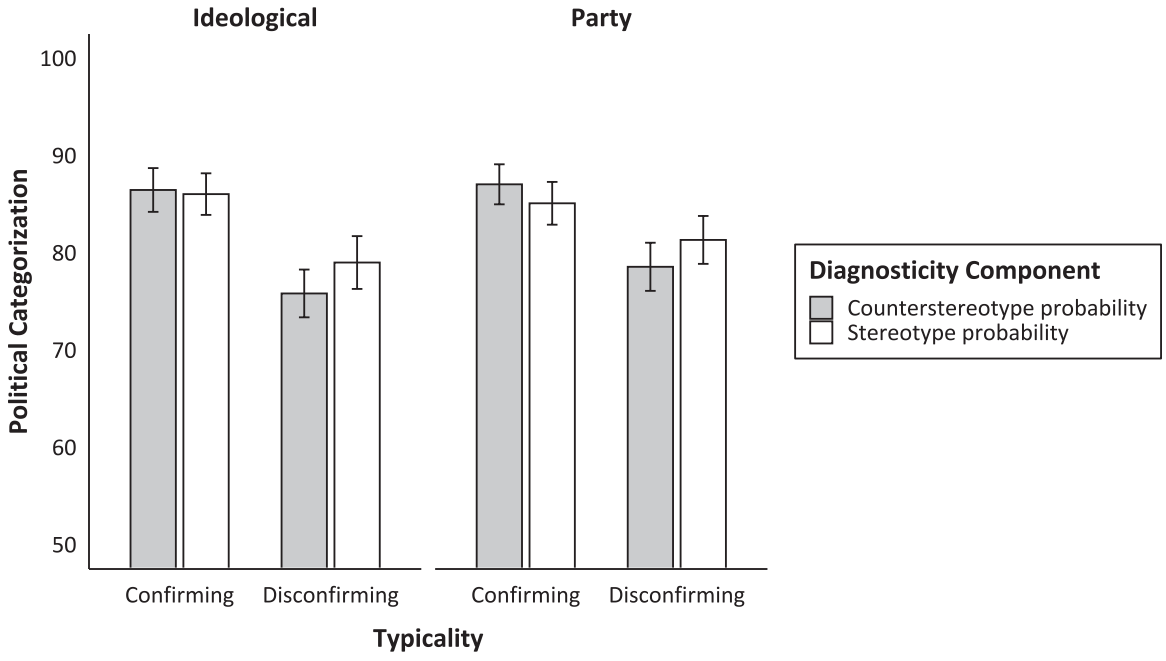


FIGURE 4 Mean categorization scores for Experiment 3. A categorization score of 100 indicates that the person with the target opinion is perceived as definitely belonging to the category associated with that opinion, while a score of 50 indicates they are seen as equally likely to belong to either category. The figure presents results after applying the outlier cutoff (i.e., after excluding values below 40). Error bars represent 95% confidence intervals.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Full model

To test whether the effect of typicality significantly differed between the category type and diagnosticity component conditions, we conducted a three-way mixed ANOVA with political categorization as the dependent variable, typicality and category type as between-subjects factors, and diagnosticity component as a within-subjects factor. We obtained a significant effect of typicality, $F(1, 867) = 66.42$, $p < .001$, $\eta_p^2 = .071$, 90% CI [.043, .095] and of diagnosticity component, $F(1, 867) = 5.38$, $p = .021$, $\eta_p^2 = .006$, 90% CI [.000, .017]. These effects were qualified by a significant interaction, $F(1, 867) = 13.48$, $p < .001$, $\eta_p^2 = .015$, 90% CI [.004, .030], supporting our conclusion that the effect of changing the counterstereotype probability is indeed larger than that of changing the stereotype probability—at least when the baseline diagnosticity is high (see Figure 5). The interaction between category type and typicality was not significant, $F(1, 867) = 2.50$, $p = .114$, $\eta_p^2 = .003$, 90% CI [.000, .011]. Thus, even with the combined sample, we could find no evidence of differential effectiveness of counterstereotypes for the two category types. All other effects were non-significant.

Expected effect in the stereotype probability condition

The smaller effect of typicality in the stereotype probability condition is not surprising: Based on the diagnosticity formula, manipulating the stereotype probability *should* have a smaller effect on political categorization than manipulating the counterstereotype probability—at least for highly diagnostic cues. To quantify this difference, we ran a simulation estimating the expected effects of manipulating the stereotype probability given the observed effects of manipulating the counterstereotype probability. That is, we identified the baseline levels of the two diagnosticity components as well as the change in the counterstereotype probability that are most likely to produce the observed effects in the counterstereotype probability condition. We

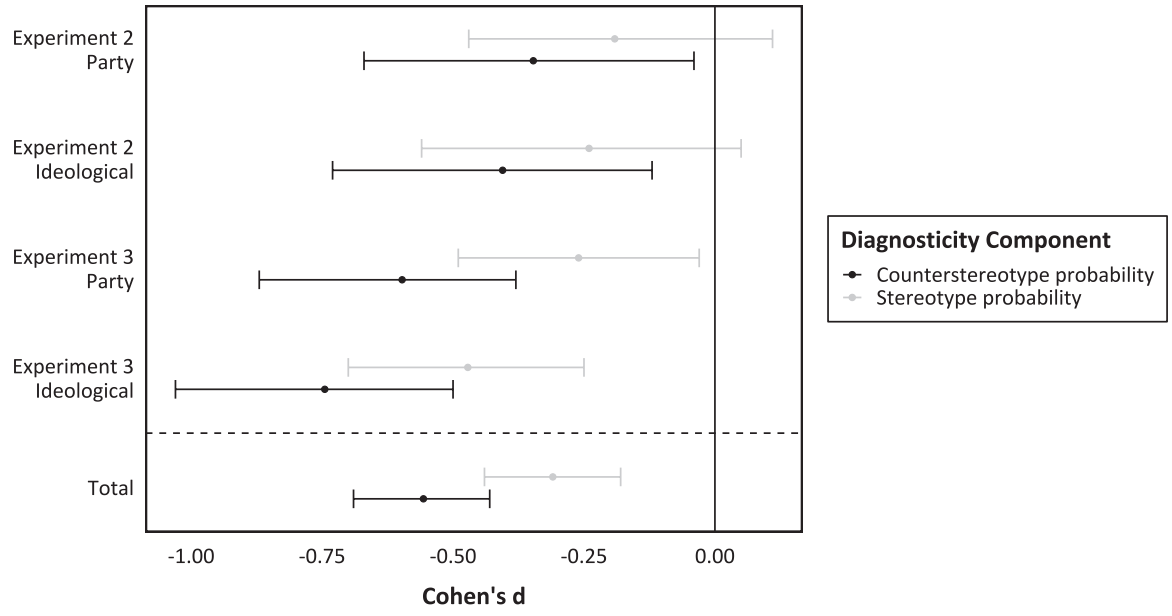


FIGURE 5 Effect sizes for Experiments 2 and 3. The forest plot presents Cohen's d for the effect of typicality, with lower values indicating greater reductions in categorization in the disconfirming compared with the confirming condition. Each row represents an independent sample (Experiment × Category Type condition). Effects of the two diagnosticity components are plotted separately. The final row shows effect size estimates calculated from the full combined dataset. Dots represent point estimates; error bars represent 95% bootstrap confidence intervals (CIs), calculated from 1000 resamples.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

then calculated the expected effect of manipulating the stereotype probability assuming the same baseline probabilities and the same magnitude of change. According to this simulation, manipulating the stereotype probability should have produced a difference between the two typicality conditions of 1.6 (for a more detailed description, see Supporting Information S9).

To test whether the effect observed in the stereotype probability condition differed from this expected effect, we conducted a two-sided independent samples *t*-test, comparing the stereotype-confirming and stereotype-disconfirming conditions, but with the null hypothesis that the observed difference (4.5) equals the expected difference (1.6) rather than zero. The test revealed that the difference between the stereotype-disconfirming condition and the stereotype-confirming condition was significantly greater than the expected difference $t(898.92) = -3.03$, $p = .003$, $d = -.20$, 95% CI [−.34, −.070]. Thus, the effects found in the stereotype probability condition were significantly larger than we would have expected if participants' political categorization followed the diagnosticity formula provided above. One potential explanation is that participants generalized what they learned about one category to the opposing category. For example, participants might have inferred from a pro-gun control Republican that opinions on guns are *generally* less correlated with partisanship than they had thought. Such generalization effects could be particularly beneficial in contexts in which replicating a potentially harmful cue—such as a racist attitude or a false belief—is undesirable. Another explanation might be that some participants responded to the categorization item using a representativeness heuristic (Kahneman & Tversky, 1973; Skowronski, 2002). Rather than following a diagnosticity-based inference process, they might have based their responses purely on how likely the cue is in the category for which it is typical (i.e., purely on the stereotype probability), thereby decreasing the observed effect in the counterstereotype probability condition.

Neutral typicality condition

To test whether the findings truly reflect a reduction in political categorization in the stereotype-disconfirming condition rather than an increase in the stereotype-confirming condition, we compared the categorization score for the target opinion with the average of the two control opinions, for which diagnosticity had not been manipulated. As these opinions had been presented as target opinions for other participants, this method controls for differences in baseline diagnosticity between the different opinions. To prevent an unbalanced design due to outlier exclusion, we randomly sampled a dataset in which the three issues appeared equally often as target and control issues. On this dataset, we conducted a two-way mixed ANOVA with political categorization as the dependent variable, typicality as a between-subjects factor, and item type (target vs. control) as a within-subjects factor (using only the data from the counterstereotype probability condition). We obtained significant effects of typicality, $F(1, 736) = 31.09$, $p < .001$, $\eta_p^2 = .041$, 90% CI [.020, .066], and item type, $F(1, 736) = 4.75$, $p = .030$, $\eta_p^2 = .006$, 90% CI [.000, .019]. These effects were qualified by a significant interaction, $F(1, 736) = 27.02$, $p < .001$, $\eta_p^2 = .035$, 90% CI [.017, .060].

Simple effects analyses revealed that the effect of item type was significant and negative in the stereotype-disconfirming condition, $t(368) = -5.55$, $p < .001$, $d_z = -.29$, 90% CI [−.37, −.20], indicating that the counterstereotypical exemplars reduced political categorization compared with the baseline level. In the stereotype-confirming condition, the effect was significant and positive, $t(368) = 2.70$, $p = .007$, $d_z = .14$, 90% CI [.050, .24], indicating that stereotypical exemplars increased political categorization compared with the baseline level. Importantly, the effect size in the stereotype-disconfirming condition was significantly larger than in the stereotype-confirming condition, $t(720.38) = 2.44$, $p = .015$, $d = .18$, 95% CI [.020, .31], suggesting the interaction was more strongly driven by the decrease in political categorization due to the counterstereotypical exemplars than by the increase due to the stereotypical ones.

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

GENERAL DISCUSSION

In polarized contexts, many behavioral cues (e.g., expressions of certain opinions) are perceived as diagnostic of political categories—likely due to increases in actual diagnosticity (driven by partisan sorting and opinion alignment) and due to exaggerated perceptions of opinion extremity and homogeneity—leading people to categorize others on the basis of these cues. The current research constitutes a proof of concept, demonstrating that category inferences based on a cue can be reduced by changing category stereotypes about that cue. Specifically, this can be achieved by exposing people to counterstereotypical exemplars, which make the perceived probabilities of observing the cue in two opposing categories more similar. In our experiments, the most effective way of reducing political categorization was presenting counterstereotypes designed to increase the perceived counterstereotype probability, that is, the probability of the cue being exhibited by members of the category for which it is less likely. Counterstereotypes designed to decrease the stereotype probability—the probability of the cue being exhibited by members of the category for which it is more likely—seem to be less effective. Thus, learning about a pro-gun control Republican is more likely to reduce categorization on the basis of pro-gun control opinions than learning about a pro-gun rights Democrat. This finding aligns with predictions from the diagnosticity formula for cases with high baseline diagnosticity. However, our results suggest that even decreasing the stereotype probability might be reasonably effective, potentially due to generalization effects.

In Experiments 2 and 3, we examined whether the effectiveness of counterstereotypes in reducing political categorization depends on the type of category. We argued that party categories, which can be defined by a simple behavioral rule and allow for rule-based categorization, may be more resistant to subtyping than more similarity-based ideological categories, from which atypical exemplars might be more easily excluded. However, we found no evidence of differential effectiveness between these two category types. Taken together, the current research suggests that for different types of political categories, categorization can be reduced by convincing perceivers that cues that are seemingly diagnostic of one category occur more frequently in the opposing category than they previously thought.

From proof of concept to real-world impact

While our research demonstrates the in-principle malleability of political categorization via counterstereotypical exemplars, the effects observed in our experiments were relatively small and potentially short-lived, as might be expected given the one-shot nature of our manipulation. In environments saturated with stereotype-confirming information, isolated interventions like ours are unlikely to shift people's categorization tendencies in enduring and meaningful ways. Effective stereotype change occurs incrementally, through repeated exposure to diverse counterstereotypical exemplars (see Weber & Crocker, 1983). Thus, we do not propose our manipulation as a ready-to-apply real-world intervention. Rather, we see the current research as a proof of concept and recommend that future studies explore ways to sustainably increase the visibility of real-life counterstereotypes in people's information environments—both as a means of developing stronger manipulations and of creating actionable interventions.

Social media platforms, in particular, illustrate the scarcity of counterstereotypes and offer a promising avenue for such efforts. These platforms, which nowadays often serve as primary sources of political news (Newman et al., 2025), are dominated by a small, highly active minority trending toward ideological extremes (Bail, 2022; Barberá & Rivero, 2015; Bor & Petersen, 2022; Pew Research Center, 2019; Robertson et al., 2024; cf. Mukerjee et al., 2022). The disproportionate visibility of these minorities may be an important cause of distorted perceptions of opinion extremity and diversity (Ahler, 2014; Dias et al., 2025; Levendusky &

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Malhotra, 2016; Robertson et al., 2024) and the subsequent overestimation of diagnosticity. Thus, a potential strategy for experimentally inducing stronger and more durable changes in perceived cue diagnosticity could be to incentivize individuals to unfollow extreme and ideologically consistent social media accounts in favor of counterstereotypical or ideologically diverse ones (see Rathje et al., 2024).

If such sustained changes to the information environment produce enduring reductions in political categorization, future research should investigate whether this also affects more distal outcomes. In polarized contexts—where political outgroups are evaluated highly negatively (Iyengar et al., 2012, 2019)—political categorization has well-documented social consequences. For example, negative evaluations of political outgroups can make people uncomfortable with discussing politics with outgroup members or having them as friends or family members (Broockman et al., 2023). Perceiving someone as a political outgroup member can also reduce trust in both the person and the information they provide (Carlin & Love, 2013; Robbett et al., 2023), and diminish empathy toward them (Cikara et al., 2011; Simas et al., 2020). To the extent that these negative outcomes arise as a consequence of political categorization, reductions in categorization should also attenuate these downstream consequences. However, empirical verification is needed.

Practical implications for reducing categorization

Translating these insights into practice requires tackling the factors behind the relative scarcity of counterstereotypical voices. One contributing factor may be a vicious cycle in which the dominance of extreme voices shapes perceptions of group norms and encourages the self-censorship of more moderate perspectives. As political discourse increasingly takes place online, traditional gatekeeping functions have weakened, allowing individuals with extreme or hostile attitudes to exert strong influence on public debate. These individuals may express harsh criticism toward counterstereotypical ingroup members for deviating from perceived group norms (black sheep effect; Marques & Paez, 1994), possibly reducing their political engagement (Bail, 2022). The resulting overrepresentation of extreme voices can lead to misperceptions of group norms (pluralistic ignorance; Prentice & Miller, 1996) or the anticipation of social conflict (Dixon et al., 2024), causing individuals to self-censor even relatively common views that are falsely perceived as non-normative (Burnett et al., 2022; Robertson et al., 2024), thereby further amplifying the overrepresentation of extreme perspectives. Breaking this cycle may require reducing the (perceived) social costs of political participation for counterstereotypical individuals, for example, through moderation on social media that reduces harassment and fosters broader participation (see Matias, 2019).

More generally, our framework suggests that policymakers and opinion leaders may want to reflect on how policies or public campaigns can inadvertently give rise to particularly potent diagnostic cues of category membership. This is especially likely when cues are highly visible and frequently encountered in daily life, such as wearing or not wearing face masks during a pandemic or adopting new norms in language use. When such visible behaviors take on strong partisan or ideological associations, they may substantially heighten the salience of political categories and their related effects.

Limitations

Although we presume that procedures in Experiments 2 and 3 reduced demand characteristics, we cannot rule out demand effects entirely. Direct measures, such as our categorization measure, are easily understood by participants, making it apparent what construct is being

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

assessed. Thus, it may still have been possible for participants to guess the goal of the research, causing them to respond in the way they think is desired. To further reduce the risk of such demand effects, future research may investigate whether similar effects can be found when using indirect measures of political categorization (Sander & Degner, 2024). Besides concealing measurement intentions, indirect measures would be beneficial because they allow to test whether counterstereotypes also reduce political categorization in *spontaneous* impression formation. While previous research suggests that political categories are also inferred spontaneously, that is, when the task does not include an explicit categorization prompt and when no category labels are made salient (Sander & Degner, 2024), it would be highly desirable to investigate in future research whether a manipulation like ours also reduces spontaneous political categorization.

Another open question concerns the meta-cognitive inclusion trigger presented in our transcripts. Across all experiments, the counterstereotypical exemplar consistently asserted that "an increasing number of" group members shared their counterstereotypical attitude. We included this statement to reduce the relative weight participants assign to their prior stereotypes, expecting this would decrease the likelihood of subtyping and increase the magnitude of stereotype change. Because we did not manipulate the inclusion of this meta-cognitive trigger, we cannot determine whether it was necessary for stereotype change to occur. However, prior research suggests that while metacognitive cues can facilitate individual-to-group generalization, they are not strictly necessary for such generalization to happen (McIntyre et al., 2016; Myers & Hvidsten, 2024). Thus, increasing the visibility of counterstereotypes should be an effective route to decreasing categorization even in the absence of such metacognitive cues. Nevertheless, subtyping may pose a barrier to real-world stereotype change, and future research should examine both the extent of this barrier and scalable solutions to address it.

An important practical limitation of our approach is that it may not be equally appropriate for all types of diagnostic cues. While our findings suggest that promoting counterstereotypical content can reduce the diagnosticity of an opinion for political categories, this strategy could be problematic when the counterstereotypical opinion in question is itself harmful—such as the endorsement of demonstrably false factual claims (Rekker & Harteveld, 2024) or expressions of prejudice. In such cases, increasing the visibility of counterstereotypical exemplars who hold these views might inadvertently legitimize them or broaden their reach. However, despite these potential adverse effects, reducing the diagnosticity of false or harmful beliefs may be a critical first step toward enabling effective belief correction: When two opposing beliefs—one true, one false—are highly diagnostic of partisanship, attempts to correct the false belief may be dismissed because communicators of the true belief are perceived as outgroup members and thus distrusted (Reinero et al., 2023). To nevertheless mitigate this risk, interventions could instead focus on highlighting counterstereotypical voices from within the group typically associated with the harmful belief (i.e., individuals who reject this belief). This approach could still reduce diagnosticity while avoiding the amplification of harmful content. Nevertheless, further empirical research is needed to assess in which contexts such interventions are beneficial, and to weigh potential costs against the intended benefits.

Finally, a further practical limitation concerns how well the counterstereotypical opinions match the diagnostic opinions they are intended to counter. For purposes of experimental control, we employed relatively strong and simplified diagnostic opinions (e.g., "The federal government should make it more difficult to buy a gun") and presented closely matched counterstereotypical ones (e.g., a Republican saying, "I'm not sure whether banning assault weapons will have a big effect, but I think we should definitely try it"). In reality, counterstereotypical opinions may sometimes be expressed with greater nuance, which could limit their ability to increase the counterstereotype probability and thereby reduce their effect on the diagnosticity of more extreme formulations. For example, if counterstereotypical Republicans only expressed support for gun control in highly nuanced or qualified ways, stronger or more

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

simplified pro-gun control opinions might still be perceived as highly diagnostic of being a Democrat. We would expect that the effects of nuanced counterstereotypical opinions would at least partially generalize to simpler or more extreme formulations, but this assumption needs to be tested in future research. However, even if counterstereotypical opinions are indeed expressed with unusual nuance or moderation and their effects do not generalize, they could still reduce diagnosticity by lowering the stereotype probability.

CONCLUSION

The present research demonstrates that political categorization based on a political opinion can be reduced by decreasing that opinion's diagnosticity. We showed that this can be achieved by presenting counterstereotypical exemplars—individuals who exhibit the opinion despite belonging to a category for which it is perceived as atypical. Because categorization can have negative consequences in polarized contexts, increasing the visibility of individuals likely to hold cross-cutting opinions may help ease polarization and promote understanding across group lines, though these downstream effects remain to be confirmed. For this purpose, future research should develop stronger, more reliable manipulations and explore strategies for amplifying the visibility of real-life counterstereotypes.

ACKNOWLEDGMENTS

This research was supported by a doctoral scholarship from the Studienstiftung des Deutschen Volkes awarded to Carsten W. Sander. We thank Jimmy Calanchini for providing valuable feedback on tailoring our materials to a US American audience. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at https://osf.io/vhmgt.

TRANSPARENCY STATEMENT

Before data collection, all hypotheses, methods, and analyses were preregistered with the Open Science Foundation (https://osf.io/vhmgt), where all materials, methods, data, and analysis scripts are publicly available. The preregistration adheres to the disclosure requirements maintained by the Center for Open Science. Artificial intelligence (ChatGPT) was used to improve language clarity and phrasing. AI did not contribute to the intellectual content, analysis, or conclusions of the work. All AI-generated suggestions were reviewed, edited, and approved by the authors.

ORCID

Carsten W. Sander https://orcid.org/0000-0003-4032-1338
Juliane Degner https://orcid.org/0000-0002-6634-8099

REFERENCES

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *Journal of Politics, 76*(3), 607–620. https://doi.org/10.1017/S0022381614000085

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics, 80*(3), 964–981. https://doi.org/10.1086/697253

Bail, C. (2022). *Breaking the social media prism*. Princeton University Press. https://doi.org/10.1515/9780691246499

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology, 114*(2), 408–446. https://doi.org/10.1086/590649

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review, 33*(6), 712–729. https://doi.org/10.1177/0894439314558836

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review, 116*(1), 1–18. https://doi.org/10.1017/S0003055421000885

Broockman, D. E., Kalla, J. L., & Westwood, S. J. (2023). Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science, 67*(3), 808–828. https://doi.org/10.1111/ajps.12719

Burnett, A., Knighton, D., & Wilson, C. (2022). The self-censoring majority: How political identity and ideology impacts willingness to self-censor and fear of isolation in the United States. *Social Media + Society, 8*(3), 20563051221123031. https://doi.org/10.1177/20563051221123031

Carlin, R. E., & Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior, 35*(1), 43–63. https://doi.org/10.1007/s11109-011-9181-x

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 216–226. https://doi.org/10.1037/0278-7393.30.1.216

Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science, 20*(3), 149–153. https://doi.org/10.1177/0963721411408713

Colombatto, C., Uddenberg, S., & Scholl, B. J. (2021). The efficiency of demography in face perception. *Attention, Perception, & Psychophysics, 83*(8), 3104–3117. https://doi.org/10.3758/s13414-021-02351-9

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Converse, P. E. (1964/2006). The nature of belief systems in mass publics (1964). *Critical Review, 18*(1–3), 1–74. https://doi.org/10.1080/08913810608443650

Degner, J. (2025). Revisiting spontaneous trait inferences: Questions, challenges, and intradisciplinary connections. *Social and Personality Psychology Compass, 19*(7), e70072. https://doi.org/10.1111/spc3.70072

Dias, N. C., Lelkes, Y., & Pearl, J. (2025). American partisans vastly under-estimate the diversity of other partisans' policy attitudes. *Political Science Research and Methods, 13*(3), 725–735. https://doi.org/10.1017/psrm.2024.36

Dixon, G. N., Lerner, B., & Bashian, S. (2024). Challenges to correcting pluralistic ignorance: False consensus effects, competing information environments, and anticipated social conflict. *Human Communication Research, 50*(3), 419–429. https://doi.org/10.1093/hcr/hqae001

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences, 22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Erickson, M., & Kruschke, J. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*(2), 107–140. https://doi.org/10.1037//0096-3445.127.2.107

Essien, I., Stelter, M., Rohmann, A., & Degner, J. (2022). Beyond stereotypes: Prejudice as an important missing force explaining group disparities. *Behavioral and Brain Sciences, 45,* e74. https://doi.org/10.1017/s0140525x21000832

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernbach, P. M., & Van Boven, L. (2022). False polarization: Cognitive mechanisms and potential solutions. *Current Opinion in Psychology, 43,* 1–6. https://doi.org/10.1016/j.copsyc.2021.06.005

Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review, 30*(4), 1273–1293. https://doi.org/10.3758/s13423-023-02252-y

Hewstone, M., Macrae, C., Griffiths, R., Milne, A., & Brown, R. (1994). Cognitive models of stereotype change (5): Measurement, development, and consequences of subtyping. *Journal of Experimental Social Psychology, 30*(6), 505–526. https://doi.org/10.1006/jesp.1994.1024

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly, 76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. https://doi.org/10.1037/h0034747

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests* (Version 0.7.2) [Computer software]. https://CRAN.R-project.org/package=rstatix

Kite, M. E., Whitley, B. E., Jr., & Wagner, L. S. (2022). *Psychology of prejudice and discrimination*. Taylor & Francis. https://doi.org/10.4324/9780367809218

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology, 68*(4), 565–579. https://doi.org/10.1037/0022-3514.68.4.565

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kunda, Z., & Oleson, K. C. (1997). When exceptions prove the rule: How extremity of deviance determines the impact of deviant examples on stereotypes. *Journal of Personality and Social Psychology, 72*(5), 965–979. https://doi.org/10.1037/0022-3514.72.5.965

Lakens, D., Pahlke, F., & Wassmer, G. (2021). *Group sequential designs: A tutorial* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/x4azm

Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 376*(1822), 20200143. https://doi.org/10.1098/rstb.2020.0143

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly, 80*(S1), 378–391. https://doi.org/10.1093/poq/nfv045

Lüders, A., Carpentras, D., & Quayle, M. (2023). Attitude networks as intergroup realities: Using network-modelling to research attitude-identity relationships in polarized political contexts. *British Journal of Social Psychology, 63,* 37–51. https://doi.org/10.1111/bjso.12665

Mandalaywala, T. M., & Legaspi, J. K. (2023). Automatic encoding across social categories in American children and adults. *Developmental Psychology, 59*(12), 2296–2303. https://doi.org/10.1037/dev0001578

Marques, J. M., & Paez, D. (1994). The 'black sheep effect': Social categorization, rejection of ingroup deviates, and perception of group variability. *European Review of Social Psychology, 5*(1), 37–68. https://doi.org/10.1080/14792779543000011

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science, 59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology, 39,* 257–277. https://doi.org/10.1111/pops.12485

Mastro, D., & Tukachinsky, R. (2011). The influence of exemplar versus prototype-based media primes on racial/ethnic evaluations. *Journal of Communication, 61*(5), 916–937. https://doi.org/10.1111/j.1460-2466.2011.01587.x

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences, 116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

McIntyre, K., Paolini, S., & Hewstone, M. (2016). Changing people's views of outgroups through individual-to-group generalisation: Meta-analytic reviews and theoretical considerations. *European Review of Social Psychology, 27*(1), 63–115. https://doi.org/10.1080/10463283.2016.1201893

Mukerjee, S., Jaidka, K., & Lelkes, Y. (2022). The political landscape of the U.S. twitterverse. *Political Communication, 39*(5), 565–588. https://doi.org/10.1080/10584609.2022.2075061

Myers, C. D. (2023). Issues, groups, or idiots? Comparing theories of partisan stereotypes. *Public Opinion Quarterly, 87*(3), 635–661. https://doi.org/10.1093/poq/nfad038

Myers, C. D., & Hvidsten, T. (2024). Who is the party? The effect of counter-stereotypical partisan exemplars on inter-party stereotypes and affect. *Political Behavior, 47,* 1317–1338. https://doi.org/10.1007/s11109-024-09991-1

Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., & Fletcher, R. (2025). *Reuters Insitute digital news report 2025*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-8QQF-JT36

Park, B. (2018). How are we apart? Continuity and change in the structure of ideological disagreement in the American public, 1980–2012. *Social Forces, 96*(4), 1757–1784. https://doi.org/10.1093/sf/sox093

Pew Research Center. (2014). *Political polarization in the American public*. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Pew Research Center. (2019). *National politics on Twitter: Small share of U.S. adults produce majority of tweets*. https://www.pewresearch.org/politics/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/

Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *28,* pp. 161–209). Academic Press. https://doi.org/10.1016/S0065-2601(08)60238-5

R Core Team. (2023). *R: A language and environment for statistical computing [Computer software]*. R Foundation for Statistical Computing.

Ramasubramanian, S. (2011). The impact of stereotypical versus counterstereotypical media exemplars on racial attitudes, causal attributions, and support for affirmative action. *Communication Research, 38*(4), 497–516. https://doi.org/10.1177/0093650210384854

Rathje, S., Pretus, C., He, J., Harjani, T., Roozenbeek, J., Gray, K., van der Linden, S., & Van Bavel, J. (2024). Unfollowing hyperpartisan social media influencers durably reduces out-party animosity. [Preprint]. PsyArXiv https://doi.org/10.31234/osf.io/acbwg

Reinero, D. A., Harris, E. A., Rathje, S., Duke, A., & Van Bavel, J. J. (2023). *Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims*. [Preprint]. PsyArXiv https://doi.org/10.31234/osf.io/z4df3

Rekker, R., & Harteveld, E. (2024). Understanding factual belief polarization: The role of trust, political sophistication, and affective polarization. *Acta Politica, 59*(3), 643–670. https://doi.org/10.1057/s41269-022-00265-4

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Robbett, A., Colón, L., & Matthews, P. H. (2023). Partisan political beliefs and social learning. *Journal of Public Economics, 220,* 104834. https://doi.org/10.1016/j.jpubeco.2023.104834

Robertson, C. E., del Rosario, K. S., & Van Bavel, J. J. (2024). Inside the funhouse mirror factory: How social media distorts perceptions of norms. *Current Opinion in Psychology, 60,* 101918. https://doi.org/10.1016/j.copsyc.2024.101918

Rothschild, J. E., Howat, A. J., Shafranek, R. M., & Busby, E. C. (2019). Pigeonholing partisans: Stereotypes of party supporters and partisan polarization. *Political Behavior, 41*(2), 423–443. https://doi.org/10.1007/s11109-018-9457-5

Sander, C. W., & Degner, J. (2024). Investigating mechanisms of political polarization: Perceivers spontaneously infer ideological categories from other people's behavior. *Journal of Applied Social Psychology, 54*(1), 5–17. https://doi.org/10.1111/jasp.13011

Simas, E. N., Clifford, S., & Kirkland, J. H. (2020). How empathic concern fuels political polarization. *American Political Science Review, 114*(1), 258–269. https://doi.org/10.1017/S0003055419000534

Skowronski, J. J. (2002). Honesty and intelligence judgments of individuals and groups: The effects of entity-related behavior diagnosticity and implicit theories. *Social Cognition, 20*(2), 136–169. https://doi.org/10.1521/soco.20.2.136.20993

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52,* 689–699. https://doi.org/10.1037/0022-3514.52.4.689

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22*(4), 377–386. https://doi.org/10.3758/BF03200864

Spicer, J., & Sanborn, A. (2017). A rational approach to stereotype change. *Proceedings of the Annual Meeting of the Cognitive Science Society, 39,* 1102–1107. https://escholarship.org/uc/item/5t407851

Tajfel, H., & Turner, J. C. (1979). An integrative theory of inter-group conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of inter-group relations* (pp. 33–47). Brooks/Cole.

Vegetti, F., & Širinić, D. (2019). Left–right categorization and perceptions of party ideologies. *Political Behavior, 41*(1), 257–280. https://doi.org/10.1007/s11109-018-9451-y

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology, 45*(5), 961–977. https://doi.org/10.1037/0022-3514.45.5.961

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sander, C. W., & Degner, J. (2026). From partisans to individuals: Lowering an opinion's diagnosticity via counterstereotypes can reduce categorization based on that opinion. *Political Psychology*, *47*, e70130. https://doi.org/10.1111/pops.70130

14679221, 2026, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/pops.70130 by Universitaet Hamburg, Wiley Online Library on [28/09/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

CHAPTER 4

Constructing the Enemy: Noun and Generic Party Labels Predict Partisan Prejudice in Social Media Comments

Sander, C. W., & Degner, J. (2026). *Constructing the enemy: Noun and generic party labels predict partisan prejudice in social media comments* [Manuscript submitted for publication]. Department of Social Psychology, University of Hamburg.

Abstract

Research suggests that affective polarization stems from social categorization processes: People sort themselves and others into political categories and perceive each other as either ingroup or outgroup members. This categorization forms the basis for intergroup conflict, producing mutual dislike and distrust between groups. However, the consequences of categorization are not uniform but scale with a group's perceived entitativity—the degree to which it is perceived as coherent and unified. Research on novel groups suggests that perceived entitativity is driven by linguistic labeling. Using fine-tuned transformer models to automatically classify Facebook posts (N = 45,780) and comments (N = 769,684), we show that this effect also holds for established political groups. Specifically, posts containing noun labels (e.g., "Senate Republicans") rather than adjective labels (e.g., "Republican senators") and using these labels generically (e.g., "Democrats want") rather than non-generically (e.g., "some Democrats want"), have higher proportions of comments expressing partisan prejudice.

Keywords: affective polarization, social media, entitativity, linguistic form, generics

Constructing the Enemy: Noun and Generic Party Labels Predict Partisan Prejudice in Social Media Comments

High levels of affective polarization—the tendency to view political outgroups more negatively than political ingroups—have become a pressing concern due to links with adverse democratic outcomes (Janssen & Turkenburg, 2025; Kingzette et al., 2021; but see Broockman et al., 2023). Much research has focused on exaggerated group stereotypes as potential drivers of these evaluative differences (e.g., Ahler & Sood, 2018; Myers, 2023; Rothschild et al., 2019). Yet decades of social psychological research demonstrate that even groups with minimal meaning or no associated stereotypes elicit systematic biases (Dunham, 2018). This suggests that affective polarization may arise not only from distorted beliefs about the opposing party but also from the basic act of categorizing people into political ingroups and outgroups (Iyengar et al., 2019).

Importantly, such categorization-based biases tend to be stronger when groups are perceived as coherent—or "entitative"—rather than as loose collections of individuals. High levels of affective polarization may therefore reflect heightened perceptions of group entitativity, raising the question of what shapes views of political groups as coherent entities. One particularly ubiquitous influence may be the way political groups are referenced in everyday language. The use of linguistic labels— especially the generic (e.g., "Democrats want") vs. non-generic use (e.g., "some Democrats want") of noun (e.g., "Senate Republicans") vs. adjective labels (e.g., "Republican senators")—suggests that group members are sufficiently alike to be treated as a single, unified collective. Experimental studies likewise find that linguistic labeling elevates perceived entitativity (e.g., Gelman et al., 2010). However, this research has largely examined novel groups in laboratory settings, leaving open how labeling functions in naturalistic political discourse.

The current research addresses this gap using large-scale analyses of social media posts and comments, testing whether the effects of linguistic labeling extend to long-standing political groups of US Republicans and Democrats. By examining real-world communication about political groups,

our work brings the research on labeling into applied contexts and sheds light on how everyday language use may contribute to affective polarization.

Affective Polarization

The term "affective polarization" refers to a state of society characterized by high levels of partisan prejudice, i.e., a state where citizens feel animosity toward their political outgroup (compared to their political ingroup). Initially developed to describe polarization within the two-party system of the United States (Iyengar et al., 2012, 2019), the concept has since been applied to various other countries including many with multiparty systems. This revealed that although the United States are characterized by a remarkable increase in affective polarization over the last decades (Iyengar et al., 2012), absolute levels of affective polarization are rather average in international comparison (Reiljan, 2020; Wagner, 2021). Indeed, high levels of affective polarization appear to be widespread and have been linked to negative outcomes, including democratic backsliding (Orhan, 2022) and declining institutional and social trust (Torcal & Thomson, 2023). At the individual level, identifying someone as a political outgroup member in an affectively polarized context has been shown to reduce empathy, trust, and willingness to engage (Broockman et al., 2023; Carlin & Love, 2013; Simas et al., 2020). Therefore, it is paramount to understand the causes of affective polarization.

Entitativity Scales Affective Polarization

Affective polarization may, at least in part, be a mere consequence of people's tendency to sort themselves and others into political in- and outgroups (Dunham, 2018). But the negative consequences categorizing someone as an outgroup member are far from uniform. Rather, they depend on the outgroup's perceived entitativity—the extent to which it "is perceived as a meaningful, unified and coherent group, as opposed to a loose set of individuals" (Beukeboom & Burgers, 2019, p. 9). Perceived entitativity intensifies categorization-based biases, increasing stereotyping (Rydell et al., 2007; Spencer-Rodgers et al., 2007), raising the likelihood that stable dispositions inferred from group members' behaviors are generalized to the whole group (Crawford

et al., 2002), polarizing warmth judgments (making warm groups seem warmer and cold groups colder) and boosting competence judgments across the board (Callahan & Ledgerwood, 2016; Dang, Liu, Ren, & Su, 2018), and making outgroups and their actions appear more threatening, particularly when the functional relation between ingroup and outgroup is conflictive (Abelson et al., 1998; Callahan & Ledgerwood, 2016; Castano, Sacchi, et al., 2003; Dang, Liu, Ren, & Gu, 2018). Moreover, elevated perceptions of outgroup entitativity are linked to a range of hostile intergroup behaviors, such as stronger discrimination in the allocation of negative outcomes (Moscatelli & Rubini, 2013), greater support for severe punishments of perpetrator groups (Newheiser et al., 2012), and increased satisfaction from displaced revenge—retaliating against a group member for harm inflicted by another member (Sjöström & Gollwitzer, 2015). They are also associated with reduced empathy for outgroup members and, at times, even pleasure in response to their suffering (Cikara et al., 2014). Lastly, entitativity also drives prejudice (Andreychik & Gill, 2015; see Agadullina & Lovakov, 2018 for a meta-analysis), with recent evidence even demonstrating this effect for US partisans (Hoyt et al., 2024). Taken together, many categorization-based biases scale with group entitativity. When individuals perceive each other as members of entitative outgroups, they stereotype, dislike, and distrust one another.

What Drives Perceived Entitativity?

Because political categories are very large and not directly observable, people cannot understand them solely through personal experience. Instead, they rely on communicated information. A news article stating that "*Republicans coordinate closely to achieve shared goals*" may, for example, increase perceived entitativity, whereas highlighting *"Democratic infighting*" may decrease it. But beyond content, entitativity can also be communicated indirectly through the form of expression. Specifically, the perceived entitativity of a category has been shown to increase by (a) referring to it with a noun rather than an adjective (e.g., "Senate Republicans" vs. "Republican senators"), and (b) talking about it using generic language (e.g., "Democrats say" vs. "Some

Democrats say"; Burgers & Beukeboom, 2020; Carnaghi et al., 2008; Gelman & Heyman, 1999; Gelman et al., 2010; Walton & Banaji, 2004).

Linguistic Form

Labeling itself heightens categorical perception: Foroni and Rothbart (2011, 2013) presented participants with silhouettes of different body types that were either labeled (e.g., anorexic, normal, obese) or left unlabeled. They found that labeling increased perceived similarity among members of the same category—a key antecedent of entitativity (Campbell, 1958; Lickel et al., 2000; Blanchard et al., 2020)—and decreased perceived similarity between categories. However, not all labels exerted equal influence: Labels such as *below average*, *average*, and *above average* produced noticeably smaller effects (Foroni & Rothbart, 2011). Similarly, the linguistic form—whether labels are nouns or adjectives—matters substantially for perceived entitativity, with noun labels leading to stronger stereotyping and reduced categorization into alternative groups (Carnaghi et al., 2008). Graf et al. (2013, 2020) extended these findings by demonstrating that linguistic form also predicts downstream consequences of entitativity, particularly outgroup prejudice. In one study, participants read a news report about a single minority group member described either with an adjective or a noun. Despite identical content, noun labels elicited greater prejudice toward the minority group, regardless of whether the individual was portrayed positively or negatively. Similarly, research on person-first language (e.g., "person with addiction") has found reductions in prejudice and related outcomes compared to noun labels (e.g., "addicts") across different group domains (e.g., Baker et al., 2022; Jackl, 2023; Kelly & Westerhoff, 2010; Lowe & Willis, 2020, 2022; but see Hoffman et al., 2025). Although no studies have directly measured entitativity as a function of linguistic form, the converging evidence indicates that noun labels—as compared to adjective labels—encourage perceivers to view groups as more coherent entities.

Genericity

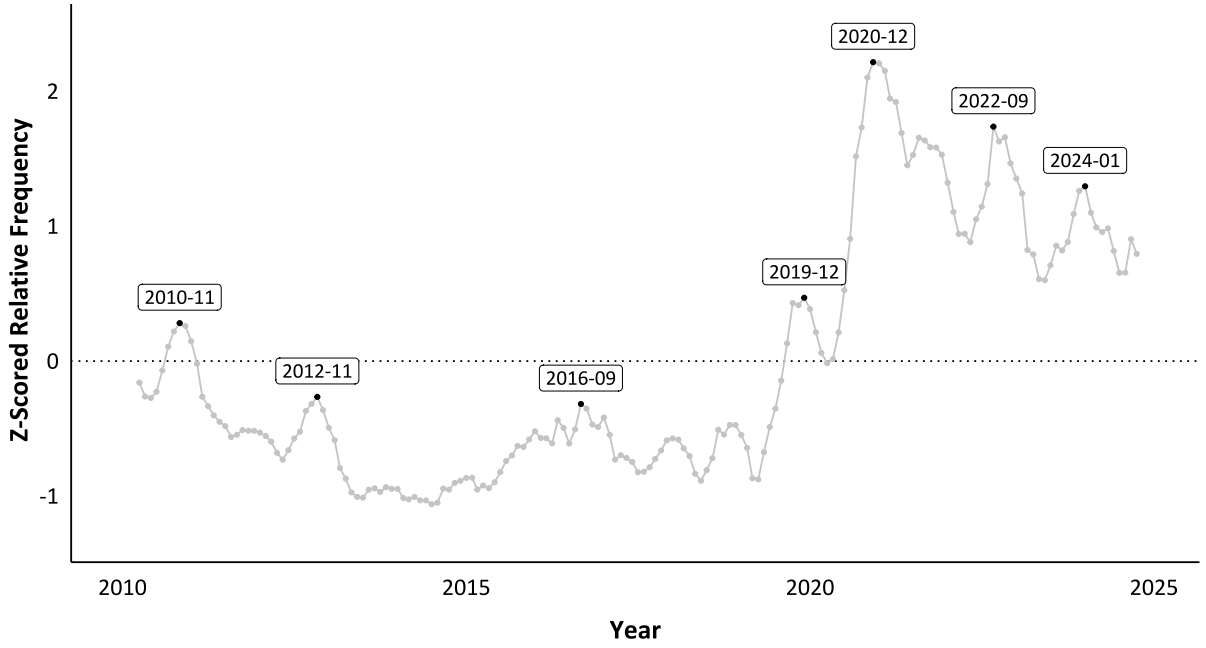
Another important aspect of labeling is whether labels generalize across group members. Statements making a general claim about a group as a whole are called *generics* and often

characterized by specific linguistic features, such as the use of bare plural nouns and the simple present tense (e.g., “Conservatives believe…”; Beukeboom & Burgers, 2019). Statements referring to only part of the group or ascribing singular actions rather than stable dispositions are non-generic (e.g., “Some conservatives said…”). Research has shown that generic expressions drive perceived entitativity. For instance, Burgers and Beukeboom (2020) presented people with statements about a fictional group of co-workers and found higher perceived entitativity when their behavior was described in a generic vs. non-generic way. Gelman et al. (2010) presented children and adults with information about a fictional animal category in generic (e.g., “Zarpies like ice cream”) or non-generic (e.g., “This Zarpie likes ice cream”) language. They found that participants who had learned generic information about the category were more likely to generalize category attributes to novel exemplars and to expect category members to share other previously unmentioned attributes (see also Hoicka et al., 2021; Noyes & Keil, 2019).

Although prior research suggests that the use of noun vs. adjective labels in generic vs. non-generic expressions increases perceived group entitativity, this research has largely examined labeling effects in laboratory settings using novel or less familiar groups. This leaves open the question of whether these effects—and their downstream consequences—generalize to real-world contexts involving well-known and highly salient political groups. Given that perceptions of these groups may already be firmly established, it is unclear whether exposure to statements about what Republicans and Democrats purportedly say, believe, or want can meaningfully shift these perceptions. Whether labeling effects replicate in this context is particularly relevant given that partisan group labels in U.S. online news have increased substantially since 2019 and remain elevated (see Figure 1), despite their known adverse effects.

Figure 1

6-Month Moving Average of the Relative Frequency of Party Labels



Note. For this plot we calculated a 6-month centered moving average of the z-scored ratio of partisan labels ("Democrat", "Democrats", "Republican", "Republicans") to non-partisan political terms ("Congress", "Legislation", "Policy", "Government", "Representative", "Senate", "White House", "President", "Administration") in the Corpus of News on the Web (NOW; Davies, 2016). The NOW corpus contains 23.4 billion words from English-language web-based newspapers and magazines from 2010 to the present. We retrieved frequencies only from U.S. sources. To ensure that changes in partisan-label frequency were not driven by shifts in corpus composition or the volume of political reporting, we normalized partisan-label counts by the combined frequency of the non-partisan terms. All frequencies reflect occurrences in which the search terms functioned as nouns.

The Current Research

The present research bridges the gap between controlled laboratory studies and real-world political discourse by testing whether labeling shapes expressions of prejudice toward long-standing political groups (i.e., U.S. Republicans and Democrats) in social media comments. Using supervised machine learning for large-scale text analysis, we examined whether linguistic labeling in Facebook posts predicted the expression of partisan prejudice in the corresponding comments. Specifically, we tested the following hypotheses:

H1: We expect a significant effect of linguistic form on comment-level partisan prejudice, with higher levels of prejudice for noun labels than for adjective labels.

H2: We expect a significant effect of genericity on comment-level partisan prejudice, with higher levels of prejudice for generic than for non-generic label use.

The current research provides an important test of the practical relevance of linguistic labeling, assessing whether patterns of partisan label use map onto expressions of partisan prejudice. To our knowledge, this constitutes the first examination of such labeling effects outside of the lab.

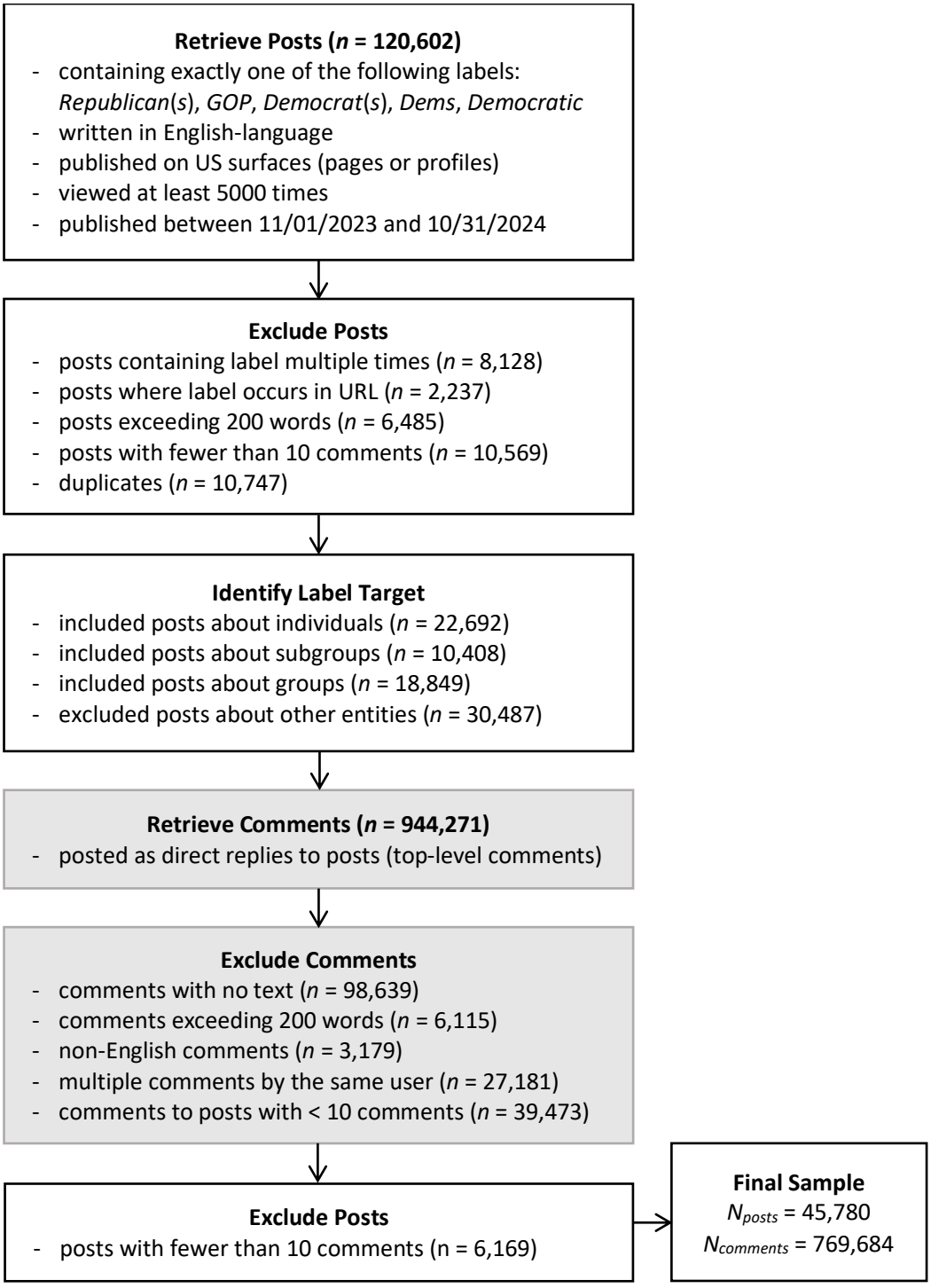
Methods

Data Collection

We tested our hypotheses using Facebook posts and comments. We selected Facebook because it is the most actively used social media platform in the U.S. with a demographically diverse user base (Pew Research Center, 2024; Statista, 2025), because of its abundant political content (including posts from news media outlets, politicians, public figures, etc.), and because its text-based post-comment structure is well-suited for testing our hypotheses. We accessed the posts and comments via the Meta Content Library API (Meta Platforms, Inc., 2024). We employed a large-scale convenience sample, exhaustively downloading all posts meeting our preselection criteria within a specified timeframe, such that sample size was determined by data availability rather than a priori power considerations. Given the volume of data obtained, the study should be well-powered to detect even small effects. Preselection criteria and data exclusions are displayed in Figure 2. A more detailed description and justification as well as information on data preprocessing can be found in online supplement S1. Our final sample consisted of N = 45,780 posts and N = 769,684 comments.

Figure 2

Flowchart of Data Collection and Exclusion Steps



Note. The flow chart illustrates the data collection process for Facebook posts and comments.

Classifiers

To efficiently analyze these posts and comments, we fine-tuned five machine-learning models. These served to (a) detect whether labels were actually used as group labels that refer to people rather than institutions or events (*label target* classifier), (b) to assess our two independent variables (*linguistic form* and *genericity* classifiers), (c) to assess our main dependent variable (*partisan prejudice in the comments* classifier), and (d) to control for post-level confounds (*partisan prejudice in the posts* classifier). The label target classifier was also trained to detect whether group labels refer to individuals, subgroups, or groups, which served as an additional control variable. All models were trained in Python using the *transformers* library (Wolf et al., 2020) and a *modernBERT-base* architecture (Warner et al., 2024), except for the linguistic form classifier, which was based on a simper *distilbert-base-uncased* model (Sanh et al., 2019). In Table 1 we describe the definitions of the constructs assessed by each classifier, as well as their performance metrics. A detailed description of the training data can be found in online supplement S2. The final models' confusion matrices can be found in online supplement S3.

Table 1

Classifiers Used to Analyze Posts and Comments

Classifier	Purpose	Training Data	P	R	F1
Label target	Detect whether a partisan label refers to *individuals* (e.g., "a Republican Representative"), a *subgroup* (e.g., "House Republicans" or "California Dems"), the entire *group* of voters, supporters, party members, politicians, or elected officials (e.g., "Democrats", "Republican officials"), or to something *else* (e.g., "The Democratic Party" or "the Republican primary").	4,000 posts Individuals: 995 Subgroup: 785 Group: 796 Else: 1,424	.93	.94	.93
Linguistic form	Detect whether the label functions as an *adjective* (e.g., "Democratic voters") or a *noun* (e.g., "voting Democrats").	2,711 posts Adjective: 1,109 Noun: 1,602	.99	1.00	1.00

Genericity	Detect whether the label is used in generic (e.g., "Republicans want a total abortion ban") or non-generic way (e.g., "Some Republicans have demanded a total abortion ban"). Generics cannot be identified by any single linguistic pattern (Rhodes et al., 2025) and were thus defined as statements ascribing (a) an enduring property or attitude, (b) a repeating or continuous behavior, or (c) a single but trait-implying behavior to a group or subgroup as an abstract whole.	3,000 posts Non-generic: 2,124 Generic: 876 600 synthetic posts	.83	.84	.83
Partisan prejudice in the posts	Detect partisan prejudice in the posts, defined as expressions of dislike or negativity toward one of the two major U.S. political parties or a party-aligned entity (e.g., "MAGA Republicans", "Trump", "the liberal media"; see Overgaard et al., 2024).	2,000 posts Non-prejudiced: 1,368 Prejudiced: 632 600 synthetic posts	.76	.87	.81
Partisan prejudice in the comments	Detect partisan prejudice in comments. Compared to posts, which were preselected to include explicit party references, comments mainly featured implicit ones. Assuming that these primarily target the labeled partisan group from the post, we coded them as prejudiced. Further, given our focus on entitativity-mediated partisan prejudice, we only counted negativity toward collective targets. Accordingly, comments were coded as prejudiced when they expressed negativity toward an unspecified collective entity (e.g., "Idiots!") or a specified party-aligned collective entity (e.g., "Dems are the worst").	4,000 comments Non-prejudiced: 3,095 Prejudiced: 905	.68	.73	.70

Note. Models were evaluated using precision (P), recall (R), and F1 scores for the minority class (underlined), except for the multi-class classifier (label target), for which we report macro-averaged scores (i.e., the average across all classes). Precision reflects how many of the posts that the model labeled as belonging to the minority category actually do belong to that category. Recall indicates how many of the actual minority-class posts the model successfully identified. The F1 score is the harmonic mean of precision and recall. For the genericity and partisan prejudice in the posts classifiers we augmented the training sets with 600 synthetic posts each, because in the natural data both constructs often co-occurred. Synthetic posts consisted of pairs of posts in which we varied only one construct while holding the other constant, to help the algorithm become sensitive only to each construct's defining cues. For a test showing that this resulted in a partisan prejudice classifier that was largely insensitive to genericity, see online supplement S4.

Transparency and Openness

We report our sample size determination and all data exclusions. Coding manuals, classifier training scripts, trained classifiers, data collection and analysis scripts are available at https://osf.io/jd5ub/overview?view_only=1ef453e65ac349cda5a939665c40e6a4. The Meta Content Library terms of use prohibit redistribution of data, such that we are unable to share the data used for classifier training and our analyses. Procedures and confirmatory analyses were preregistered at https://osf.io/gd4t5/overview?view_only=0f4a9640dcd2440f868ddaccd66bd3dd. All procedures were approved by [removed for peer review]. Analyses were conducted in R (R Core Team, 2025) using the glmmTMB package (Brooks et al., 2017).

Results

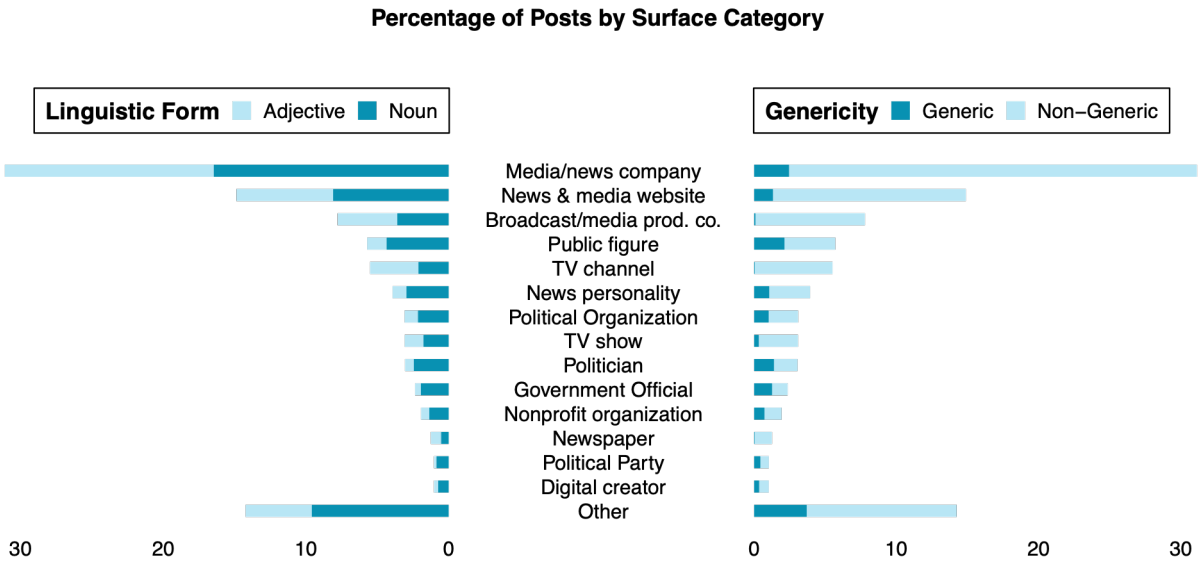
In the manuscript we only report results from the confirmatory analyses. In the online supplement we report results from additional analyses, including zero-order correlations between predictors (online supplement S5) and models testing the effect of labeling on rates of angry reactions (online supplement S6).

Description of the Data

We analyzed a total of *N* = 769,684 comments nested in *N* = 45,780 posts nested in *N* = 3,878 surfaces (99% pages and 1% profiles) and posted by *N* = 391,386 unique users. The largest share of posts in our sample came from news media outlets (see Figure 3). Despite exhibiting lower relative frequencies of generic or noun labels compared to public figures, politicians, or government officials, they still contributed many generic and especially noun label posts in absolute numbers (for differences between news media outlets see online supplement S7).

Figure 3

Percentage of Posts by Surface Category and Linguistic Form/Genericity



Note. The plot shows the percentage of posts by surface category and linguistic form (left side) or genericity (right side). Surface categories are ordered by the number of posts in our sample, with categories contributing less than 1 percent summarized in the *Other* category.

Confirmatory Analyses

To test our hypotheses, we fitted multilevel logistic regressions (using a logit link function) with random intercepts for surfaces and posts within surfaces and fixed effects for linguistic form, genericity, label target (individuals vs. subgroup vs. group), and partisan prejudice. For Hypothesis 1 we used the full dataset including posts where the label referred to individuals. Because genericity is undefined for these posts, we did not include it in this model. For Hypothesis 2 we excluded posts about individuals and included genericity as a predictor. To separate post-level from surface-level effects, we applied a multilevel decomposition to all predictors using group-mean centering to estimate within-surface effects and reintroducing the corresponding surface means to estimate between-surface effects. As surfaces may differ on unobserved confounds, we focused on within-surface effects for our hypothesis tests.

Hypothesis 1: Linguistic Form

To test H1, we fitted three models: an intercept-only baseline (Model 0), a model adding within- and between-surface linguistic form (Model 1), and a model additionally controlling for within- and between-surface label target (individual, subgroup, group) and post-level partisan prejudice (Model 2). Across Models 1 and 2, within-surface linguistic form significantly predicted comment-level partisan prejudice, with noun labels predicting a higher probability of partisan prejudice than adjective labels (see Table 2). The estimated effect for within-surface linguistic form was *OR* = 1.39, 95% CI [1.37,

Table 2

Random-Intercept Logistic Regression Models Predicting Partisan Prejudice in the Comments from Linguistic Form ($N_{comments}$ *= 769,684;* N_{posts} *= 45,780;* $N_{surfaces}$ *= 3,878;* $N_{commenters}$ *= 391,386)*

	Model 0	Model 1	Model 2
Surface-level variables			
Proportion of noun labels		1.78 [1.67, 1.90]***	1.07 [0.98, 1.16]
Proportion of prejudiced posts			1.33 [1.25, 1.42]***
Proportion of subgroup references			1.76 [1.60, 1.93]***
Proportion of group references			1.73 [1.58, 1.91]***
Post-level variables			
Linguistic form (Noun)		1.39 [1.37, 1.42]***	1.15 [1.13, 1.18]***
PP (Prejudiced)			1.17 [1.14, 1.20]***
Label target (Subgroup)			1.55 [1.51, 1.59]***
Label target (Group)			1.33 [1.29, 1.36]***
$ICC_{surfaces}$	.052	.042	.037
ICC_{posts}	.100	.096	.090
BIC	778,846	777,347	775,604
$R^2_{conditional}$	.152	.146	.147
$R^2_{marginal}$	.000	.010	.022

Note. Odds ratios (*OR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces and posts. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

1.42], $p < .001$ in Model 1 and *OR* = 1.15, 95% CI [1.13, 1.18], $p < .001$ in Model 2. The effect size decreased after including covariates, but the direction and significance of the effect remained stable, suggesting that the relationship between linguistic form and comment-level partisan prejudice was robust to these adjustments. Based on Model 2, moving from an adjective to a noun label increased the predicted probability of partisan prejudice from 14.8% to 16.7% (for a typical surface with 59% noun labels).

Hypothesis 2: Genericity

To H2, we followed the same modelling approach (see Table 3). Across Models 1 and 2, within-surface genericity significantly predicted higher levels of partisan prejudice in the comments, with generic label use predicting a higher probability of prejudiced comments non-generic labeling. The estimated effect for within-surface genericity was *OR* = 1.22, 95% CI [1.19, 1.26], $p < .001$ in Model 1 and *OR* = 1.20, 95% CI [1.17, 1.24], $p < .001$ in Model 2. The direction and magnitude of the effect remained stable, suggesting that the relationship between genericity and comment-level partisan prejudice was robust to these adjustments. Based on Model 2, moving from non-generic to generic label use increased the predicted probability of partisan prejudice from 19.2% to 22.2% (for a typical surface with 29% generic posts). The effect of within-surface linguistic form remained significant in this analysis (*OR* = 1.16, 95% CI [1.12, 1.20], $p < .001$), suggesting that genericity and linguistic form have independent contributions in explaining partisan prejudice.

Discussion

Affective polarization may already arise from the simple act of categorizing oneself and others into political ingroups and outgroups (see Dunham, 2018). Such categorization-based biases tend to be stronger when groups are seen as highly entitative. Thus, heightened affective polarization may partly reflect inflated perceptions of group entitativity. One route by which people may come to perceive groups as more entitative is through linguistic labeling—specifically, exposure to generic noun labels

(e.g., “Republicans want”) rather than non-generic adjective labels (e.g., “some Republican officials want”). The present research tested the practical implications of such effects. Using a large-scale analysis of natural language data, we examined whether linguistic labeling in Facebook posts predicts the degree of partisan prejudice expressed in the corresponding comments. Our hypotheses were confirmed: Posts had more prejudiced comments when they used noun rather than adjective labels and generic rather than non-generic expressions.

Table 3

Random-Intercept Logistic Regression Models Predicting Partisan Prejudice in the Comments from Genericity ($N_{comments}$ *= 437,744;* N_{posts} *= 26,026;* $N_{surfaces}$ *= 2,785;* $N_{commenters}$ *= 248,334)*

	Model 0	Model 1	Model 2
Surface-level variables			
Proportion of generic posts		1.13 [1.06, 1.21]***	0.88 [0.80, 0.96]**
Proportion of noun labels			1.05 [0.94, 1.17]
Proportion of prejudiced posts			1.41 [1.29, 1.53]***
Proportion of group references			0.98 [0.90, 1.07]
Post-level variables			
Genericity (Generic)		1.22 [1.19, 1.26]***	1.20 [1.17, 1.24]***
Linguistic form (Noun)			1.16 [1.12, 1.20]***
PP (Prejudiced)			1.08 [1.05, 1.12]***
Label target (Group)			0.85 [0.83, 0.88]***
$ICC_{surfaces}$	.035	.035	.032
ICC_{posts}	.089	.088	.087
BIC	480,466	480,278	480,090
$R^2_{conditional}$	.124	.124	.123
$R^2_{marginal}$	.000	.002	.005

Note. Odds ratios (*OR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces and posts. *BIC* = Bayesian information criterion calculated following Lorah and Womack’s (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

Causality

While our design is observational, several features strengthen confidence in a causal interpretation. First, our findings align with experimental research demonstrating causal effects of labeling on entitativity and prejudice (e.g., Gelman et al., 2010; Graf et al., 2013, 2020). Although prior work has also documented reverse causality—with high entitativity perceptions prompting increased labeling (Carnaghi et al., 2008; Rhodes et al., 2012)—this alternative is implausible in our context, as comments are written in response to the posts and by different users. Second, although we cannot entirely rule out unobserved post-level confounds, we addressed many potential confounders by controlling for partisan prejudice in the posts and label target (i.e., whether labels referred to individuals, subgroups, or entire groups) as well as for unobserved between-surface differences (by examining only within-surface effects). The hypothesized effects remained robust to these adjustments. Finally, supplementary moderation analyses (see online supplement S8) showed that effects were generally consistent across label target levels and for both labeled groups (Republicans and Democrats), further reducing the likelihood of spurious correlations. Taken together, these considerations provide grounds for regarding a causal interpretation as plausible, though future field experiments manipulating linguistic framing in naturalistic social media contexts could offer more definitive evidence.

Mechanism

Under a causal interpretation, we assume that the relationship between labeling and partisan prejudice is mediated by perceived outgroup entitativity. This view is in line with extant research showing that labeling creates perceptions of the outgroup as a unified collective (Burgers & Beukeboom, 2020; Carnaghi et al., 2008; Foroni & Rothbart, 2011, 2013; Gelman et al., 2010; Novoa et al., 2023) and that this perception, in turn, amplifies partisan prejudice (Hoyt et al., 2024). Further support comes from additional analyses reported in the online supplement, that can rule out three alternative mechanisms: First, one might argue that the findings are due to selection effects rather than within-user change (see Mamakos & Finkel, 2023), with posts using generic noun labels eliciting

comments disproportionately from users who are already highly prejudiced. Contrary to this idea, we could replicate our findings even when controlling for differences between commenters, suggesting that linguistic labeling affects *how* users comment rather than *which* users comment (see online supplement S9). Second, it might be argued that entitativity does not directly increase outgroup prejudice but instead amplifies the impact of partisan animosity in the post—for instance by strengthening individual-to-group generalization (Crawford et al., 2002). Moderation analyses showing that labeling effects were not stronger but even descriptively weaker for prejudiced posts, contradict this idea (see online supplement S8). Lastly, the possibility that effects are driven by ingroup rather than outgroup labeling and entitativity appears unlikely given that commenters should predominantly express prejudice toward their outgroups and our analyses show that the groups targeted by commenters' prejudicial expressions were usually the ones labeled in the posts—suggesting that from the commenters' perspective, labels were usually outgroup labels (see online supplement S10).

One additional route through which labeling may exert its influence, is through perceived essentialism—the extent to which characteristics ascribed to a group are "perceived to be immutable to its members, and stable across time and situations" (Beukeboom & Burgers, 2019, p. 10). Essentialism is positively related to entitativity (e.g., Hoyt et al., 2024) and also affected by labeling (Gelman et al., 2010; Rhodes et al., 2012, but see Hoicka et al., 2021)—though more so for plausibly biological properties than cultural ones (Noyes & Keil, 2019). Although essentialist beliefs often predict more negative intergroup outcomes (e.g., interethnic or racial prejudice; Diesendruck & Menahem, 2015; Jayaratne et al., 2006) they can also have neutral or even positive effects. For example, attributing homosexuality to a biological essence predicted greater tolerance, presumably because sexual orientation was viewed as beyond personal control (Haider-Markel & Joslyn, 2008; Jayaratne et al., 2006). For partisan prejudice specifically, findings are mixed, with both positive (Suhay et al., 2017) and negative (Hoyt et al., 2024) associations reported. While it is plausible that

labeling influences partisan prejudice in part through perceived essentialism, the available evidence makes essentialism a comparatively less likely mechanism.

A final consideration concerns the nature of our dependent variable. Previous literature investigating effects of labeling or entitativity assessed prejudice via self-reports, asking participants to express their attitudes toward outgroups (e.g., Graf et al., 2020; Hoyt et al., 2024). In contrast, our dependent measure captured a singular unprompted behavior, namely the expression of prejudice in a written comment. At the current point in time, we cannot determine whether the observed labeling effects represent enduring shifts in prejudiced attitudes or merely a temporal activation of existing attitudes. Indeed, given the ubiquity of political group labels and other entitativity-relevant information in everyday life, it seems unlikely that lasting attitudinal change would follow from a single labeling exposure. However, as we will discuss below, this would not necessarily negate the practical relevance of the observed effects.

To summarize, while alternative mechanisms could theoretically explain the observed associations, our additional analyses and theoretical considerations converge with extant research to support the proposed mediation model: Labeling heightens perceived outgroup entitativity, which in turn amplifies partisan prejudice or the expression thereof.

Implications

The potentially fleeting nature of the effects, together with the—by conventional standards—very small effect sizes, raise the question whether the observed effects are of any practical interest. However, under specific conditions even small or transient effects may have substantive real-world significance. Anvari et al. (2023) argue that small effects can gain relevance in the presence of *amplifying mechanisms* and absence of *counteracting mechanisms*. In the following we will discuss which of these mechanisms apply to our findings.

The first mechanism that may amplify the observed effects is that they may *accumulate through repetition*. Reading or hearing about political groups is not a once-in-a-lifetime event—we experience it repeatedly and often on a daily basis. Being confronted with multiple instances of

generic noun labels vs. non-generic adjective labels may slowly lead to shifts in habitual group representations. Second, given the extreme asymmetry of political communication, effects are likely to *scale up*—reaching a vast number of people. And third, there is reason to assume that labeling may trigger *cascades* of downstream consequences. For instance, given that labeling was found to be not just a consequence but also an expression of entitativity and essentialism (Carnaghi et al., 2008; Rhodes et al., 2012), perceptions of high or low entitativity resulting from one person's label use could affect another person's label use, and thus potentially spread through social networks. Furthermore, even if labeling effects are only transient and do not accumulate, the heightened public expressions of partisan prejudice we observed may nevertheless contribute to affective polarization: Comment sections have been shown to affect perceptions of public opinion (e.g., Lee et al., 2022) and could therefore make other users develop exaggerated assumptions about their outgroup's attitude toward their ingroup (negative meta-perceptions), which could in turn exacerbate affective polarization (Lees & Cikara, 2020; Moore-Berg et al., 2020; Ruggeri et al., 2021). More generally, to the extent that our findings are driven by perceived entitativity, many other negative downstream consequences are plausible, given the large number of intergroup phenomena scaled by entitativity—including stereotyping, individual-to-group generalization, threat perceptions, discrimination in negative outcome allocations, support for severe punishment, greater satisfaction from displaced revenge, or reduced empathy (Abelson et al., 1998; Callahan & Ledgerwood, 2016; Castano, Sacchi, et al., 2003; Cikara et al., 2014; Crawford et al., 2002; Dang, Liu, Ren, & Gu, 2018; Moscatelli & Rubini, 2013; Newheiser et al., 2012; Rogier & Yzerbyt, 1999; Rydell et al., 2007; Sjöström & Gollwitzer, 2015; Spencer-Rodgers et al., 2007; Yzerbyt et al., 1998).

The counteracting mechanisms described by Anvari et al. (2023) appear less likely to apply. To begin with, it seems unlikely that labeling would cease to have an effect through *habituation* as for both linguistic form and genericity, both factor levels are already highly prevalent. Thus, if habituation were possible, then it should already have taken place. For similar reasons, *counteracting responses* are rather unlikely to occur, as in most cases, non-generic adjective labels

should be too unobtrusive to provoke reactance. One counteracting mechanism that could reduce the practical relevance of our findings is *homeostasis*—the idea that a system (e.g., individuals' entitativity perceptions) reverts to a stable set-point over time. This could be the case if factors other than linguistic labeling produce much larger and/or more durable shifts in perceived entitativity, thereby reinstating a previous set-point. There is limited research on other drivers of entitativity, showing, for example, that group identity symbols such as flags or logos can increase entitativity (Callahan & Ledgerwood, 2016). However, many factors that may be more directly relevant to entitativity (e.g., reports of partisan infighting) have—to our knowledge—not yet been investigated, and no studies have compared the relative strength of different causal factors. Thus, future research is necessary to gauge the degree to which homeostasis may counteract effects of linguistic labeling. But as we have argued above, even if entitativity perceptions routinely revert back to stable set-points, there may be crystallized consequences of short-term shifts in entitativity—like prejudiced Facebook comments—that continue to exert influence even after the internal state that produced them has faded.

Taken together, our discussion of amplifying and counteracting mechanisms suggest that our findings have substantial practical relevance—despite small effect sizes. Whether directly via repetition-based accumulation, or indirectly via increasing negative meta-perceptions through expressions of partisan prejudice, linguistic labeling represents a plausible contributor to affective polarization. Given the potential adverse outcomes of affective polarization for democratic functioning and social cohesion (Broockman et al., 2023; Carlin & Love, 2013; Orhan, 2022; Simas et al., 2020; Torcal & Thomson, 2023), future research may build on our work by manipulating linguistic labeling in naturalistic contexts (e.g., in collaboration with news media outlets) and over an extended period of time to verify the causal interpretation of our results and assess the potential of labeling-based interventions. If effects prove to be causal, future translational research should conduct more fine-grained analyses of labeling variations to derive clear recommendations, for instance for journalistic praxis. This includes contrasting noun and adjective labels not only with each

other but also with person-first constructions (e.g., “people who voted Republican”) or entirely non-partisan framing (Shehata, 2014), as well as disentangling specific linguistic features that reduce genericity, such as quantifiers (e.g., “some Republicans”) versus references to discrete actions rather than enduring traits (e.g., “Republicans voted” versus “Republicans are”).

Conclusion

In sum, our findings suggest a reliable albeit small association between the use of generic vs. non-generic and noun vs. adjective labels in social media posts and increased expressions of partisan prejudice in subsequent comments. The results are consistent with experimental research proposing that linguistic labeling can lead to increased prejudice via perceived outgroup entitativity. Given that the small effects may accumulate via repetition, scale up due to the asymmetry of political communication, and may trigger cascades of downstream consequences, they may have substantial practical relevance with potential implications for affective polarization. Although the findings highlight the potential value of more careful language use in political communication, particularly within journalism and on social media, more research is needed to determine the long-term impact on partisan prejudice and broader polarization dynamics.

References

Abelson, R. P., Dasgupta, N., Park, J., & Banaji, M. R. (1998). Perceptions of the collective other. *Personality and Social Psychology Review*, *2*(4), 243–250. https://doi.org/10.1207/s15327957pspr0204_2

Agadullina, E. R., & Lovakov, A. V. (2018). Are people more prejudiced towards groups that are perceived as coherent? A meta-analysis of the relationship between out-group entitativity and prejudice. *British Journal of Social Psychology*, *57*(4), 703–731. https://doi.org/10.1111/bjso.12256

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, *80*(3), 964–981. https://doi.org/10.1086/697253

Andreychik, M. R., & Gill, M. J. (2015). Do natural kind beliefs about social groups contribute to prejudice? Distinguishing bio-somatic essentialism from bio-behavioral essentialism, and both of these from entitativity. *Group Processes & Intergroup Relations*, *18*(4), 454–474. https://doi.org/10.1177/1368430214550341

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2023). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*, *18*(2), 503–507. https://doi.org/10.1177/17456916221091565

Baker, E. A., Hamilton, M., Culpepper, D., McCune, G., & Silone, G. (2022). The effect of person-first language on attitudes toward people with addiction. *Journal of Addictions & Offender Counseling*, *43*(1), 38–49. https://doi.org/10.1002/jaoc.12102

Beukeboom, C. J., & Burgers, C. (2019). How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, *7*, 1–37. https://doi.org/10.12840/issn.2255-4165.017

Blanchard, A. L., Caudill, L. E., & Walker, L. S. (2020). Developing an entitativity measure and distinguishing it from antecedents and outcomes within online and face-to-face groups. *Group Processes & Intergroup Relations*, *23*(1), 91–108. https://doi.org/10.1177/1368430217743577

Brandt, M. J., & Vallabha, S. (2025). Intraindividual changes in political identity strength (but not direction) are associated with political animosity in the United States and the Netherlands. *Personality and Social Psychology Bulletin*, *51*(5), 828–844. https://doi.org/10.1177/01461672231203471

Broockman, D. E., Kalla, J. L., & Westwood, S. J. (2023). Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science*, *67*(3), 808–828. https://doi.org/10.1111/ajps.12719

Brooks, M., E., Kristensen, K., Benthem, K., J. ,van, Magnusson, A., Berg, C., W., Nielsen, A., Skaug, H., J., Mächler, M., & Bolker, B., M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Burgers, C., & Beukeboom, C. J. (2020). How language contributes to stereotype formation: Combined effects of label types and negation use in behavior descriptions. *Journal of Language and Social Psychology*, *39*(4), 438–456. https://doi.org/10.1177/0261927X20933320

Callahan, S. P., & Ledgerwood, A. (2016). On the psychological function of flags and logos: Group identity symbols increase perceived entitativity. *Journal of Personality and Social Psychology*, *110*(4), 528–550. https://doi.org/10.1037/pspi0000047

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*(1), 14–25.

Carlin, R. E., & Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior*, *35*(1), 43–63. https://doi.org/10.1007/s11109-011-9181-x

Carnaghi, A., Maass, A., Gresta, S., Bianchi, M., Cadinu, M., & Arcuri, L. (2008). Nomina sunt omina: On the inductive potential of nouns and adjectives in person perception. *Journal of Personality and Social Psychology*, *94*(5), 839–859. https://doi.org/10.1037/0022-3514.94.5.839

Castano, E., Sacchi, S., & Gries, P. H. (2003). The perception of the other in international relations: Evidence for the polarizing effect of entitativity. *Political Psychology*, *24*(3), 449–468. https://doi.org/10.1111/0162-895X.00336

Castano, E., Yzerbyt, V., & Bourguignon, D. (2003). We are one and I like it: The impact of ingroup entitativity on ingroup identification. *European Journal of Social Psychology*, *33*(6), 735–754. https://doi.org/10.1002/ejsp.175

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125. https://doi.org/10.1016/j.jesp.2014.06.007

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology*, *83*(5), 1076–1094. https://doi.org/10.1037/0022-3514.83.5.1076

Dang, J., Liu, L., Ren, D., & Gu, Z. (2018). "Groupy" allies are more beneficial while "groupy" enemies are more harmful. *Social Psychological and Personality Science*, *9*(8), 925–934. https://doi.org/10.1177/1948550617729409

Dang, J., Liu, L., Ren, D., & Su, Q. (2018). Polarization and positivity effects: Divergent roles of group entitativity in warmth and competence judgments. *Journal of Experimental Social Psychology*, *74*, 74–84. https://doi.org/10.1016/j.jesp.2017.09.003

Davies, M. (2016). *Corpus of news on the web (NOW)* [Dataset]. https://www.english-corpora.org/now/

Diesendruck, G., & Menahem, R. (2015). Essentialism promotes children's inter-ethnic bias. *Frontiers in Psychology*, *6*, Article 1180. https://doi.org/10.3389/fpsyg.2015.01180

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences*, *22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Foroni, F., & Rothbart, M. (2011). Category boundaries and category labels: When does a category name influence the perceived similarity of category members? *Social Cognition*, *29*(5), 547–576. https://doi.org/10.1521/soco.2011.29.5.547

Foroni, F., & Rothbart, M. (2013). Abandoning a label doesn't make it disappear: The perseverance of labeling effects. *Journal of Experimental Social Psychology*, *49*(1), 126–131. https://doi.org/10.1016/j.jesp.2012.08.002

Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science*, *10*(6), 489–493. https://doi.org/10.1111/1467-9280.00194

Gelman, S. A., Ware, E. A., & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive Psychology*, *61*(3), 273–301. https://doi.org/10.1016/j.cogpsych.2010.06.001

Graf, S., Bilewicz, M., Finell, E., & Geschke, D. (2013). Nouns cut slices: Effects of linguistic forms on intergroup bias. *Journal of Language and Social Psychology*, *32*(1), 62–83. https://doi.org/10.1177/0261927X12463209

Graf, S., Linhartova, P., & Sczesny, S. (2020). The effects of news report valence and linguistic labels on prejudice against social minorities. *Media Psychology*, *23*(2), 215–243. https://doi.org/10.1080/15213269.2019.1584571

Haider-Markel, D. P., & Joslyn, M. R. (2008). Beliefs about the origins of homosexuality and support for gay rights: An empirical test of attribution theory. *Public Opinion Quarterly*, *72*(2), 291–310. https://doi.org/10.1093/poq/nfn015

Hoffman, R. K., Post, S. M., Dodge, T., & Stock, M. L. (2025). The effect of person-first language on obesity stigma. *American Journal of Health Promotion*, *39*(3), 388–393. https://doi.org/10.1177/08901171241284551

Hoicka, E., Saul, J., Prouten, E., Whitehead, L., & Sterken, R. (2021). Language signaling high proportions and generics lead to generalizing, but not essentializing, for novel social kinds. *Cognitive Science*, *45*(11), e13051. https://doi.org/10.1111/cogs.13051

Hoyt, C. L., Burnette, J. L., & Moore, M. (2024). Partisan prejudice: The role of beliefs about the unchanging nature of ideology and partisans. *Personality and Social Psychology Bulletin*, Advance online publication. https://doi.org/10.1177/01461672241283862

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jackl, H. M. (2023). The effects of language on the stigmatization and exclusion of returning citizens: Results from a survey experiment. *Punishment & Society*, *25*(2), 471–499. https://doi.org/10.1177/14624745211059318

Janssen, L., & Turkenburg, E. (2025). Breaking free from linear assumptions: Unravelling the relationship between affective polarization and democratic support. *European Journal of Political Research*, *64*(3), 1465–1479. https://doi.org/10.1111/1475-6765.12725

Jayaratne, T. E., Ybarra, O., Sheldon, J. P., Brown, T. N., Feldbaum, M., Pfeffer, C. A., & Petty, E. M. (2006). White Americans' genetic lay theories of race differences and sexual orientation: Their relationship with prejudice toward blacks, and gay men and lesbians. *Group Processes & Intergroup Relations*, *9*(1), 77–94. https://doi.org/10.1177/1368430206059863

Kelly, J. F., & Westerhoff, C. M. (2010). Does it matter how we refer to individuals with substance-related conditions? A randomized study of two commonly used terms. *International Journal of Drug Policy*, *21*(3), 202–207. https://doi.org/10.1016/j.drugpo.2009.10.010

Kingzette, J., Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). How affective polarization undermines support for democratic norms. *Public Opinion Quarterly*, *85*(2), 663–677. https://doi.org/10.1093/poq/nfab029

Lee, S., Atkinson, L., & Sung, Y. H. (2022). Online bandwagon effects: Quantitative versus qualitative cues in online comments sections. *New Media & Society*, *24*(3), 580–599. https://doi.org/10.1177/1461444820965187

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, *4*(3), 279–286. https://doi.org/10.1038/s41562-019-0766-4

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, *78*(2), 223–246. https://doi.org/10.1037/0022-3514.78.2.223

Lorah, J., & Womack, A. (2019). Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behavior Research Methods*, *51*(1), 440–450. https://doi.org/10.3758/s13428-018-1188-3

Lowe, G. T., & Willis, G. M. (2020). "Sex offender" versus "person": The Influence of labels on willingness to volunteer with people who have sexually abused. *Sexual Abuse*, *32*(5), 591–613. https://doi.org/10.1177/1079063219841904

Lowe, G. T., & Willis, G. M. (2022). Do popular attitudinal scales perpetuate negative attitudes towards persons who have sexually offended? *Journal of Sexual Aggression*, *28*(2), 231–243. https://doi.org/10.1080/13552600.2021.2009050

Mamakos, M., & Finkel, E. J. (2023). The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus*, *2*(10), Article pgad325. https://doi.org/10.1093/pnasnexus/pgad325

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Meta Platforms, Inc. (2024). *Meta Content Library and Content Library API* [Computer software]. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.48680/meta.metacontentlibraryapi.5.0

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*, *117*(26), 14864–14872. https://doi.org/10.1073/pnas.2001263117

Moscatelli, S., & Rubini, M. (2013). The impact of group entitativity on negative outcome allocations. *The Journal of Social Psychology*, *153*(2), 149–160. https://doi.org/10.1080/00224545.2012.712068

Myers, C. D. (2023). Issues, groups, or idiots? Comparing theories of partisan stereotypes. *Public Opinion Quarterly*, *87*(3), 635–661. https://doi.org/10.1093/poq/nfad038

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Newheiser, A.-K., Sawaoka, T., & Dovidio, J. F. (2012). Why do we punish groups? High entitativity promotes moral suspicion. *Journal of Experimental Social Psychology*, *48*(4), 931–936. https://doi.org/10.1016/j.jesp.2012.02.013

Novoa, G., Echelbarger, M., Gelman, A., & Gelman, S. A. (2023). Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, *120*(47), Article e2309361120. https://doi.org/10.1073/pnas.2309361120

Noyes, A., & Keil, F. C. (2019). Generics designate kinds but not always essences. *Proceedings of the National Academy of Sciences*, *116*(41), 20354–20359. https://doi.org/10.1073/pnas.1900105116

Orhan, Y. E. (2022). The relationship between affective polarization and democratic backsliding: Comparative evidence. *Democratization*, *29*(4), 714–735. https://doi.org/10.1080/13510347.2021.2008912

Overgaard, C. S. B., Lukito, J., & Soorholtz, K. (2024). The manifestation of affective polarization on social media: A cross-platform supervised machine learning approach. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 1160–1173. https://doi.org/10.1609/icwsm.v18i1.31380

Pew Research Center. (2024). *Social media fact sheet*. https://www.pewresearch.org/internet/fact-sheet/social-media/

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reiljan, A. (2020). 'Fear and loathing across party lines' (also) in Europe: Affective polarisation in European party systems. *European Journal of Political Research*, *59*(2), 376–396. https://doi.org/10.1111/1475-6765.12351

Rhodes, M., Gelman, S. A., & Leslie, S.-J. (2025). How generic language shapes the development of social thought. *Trends in Cognitive Sciences*, *29*(2), 122–132. https://doi.org/10.1016/j.tics.2024.09.012

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531. https://doi.org/10.1073/pnas.1208951109

Rogier, A., & Yzerbyt, V. (1999). Social attribution, correspondence bias, and the emergence of stereotypes. *Swiss Journal of Psychology / Schweizerische Zeitschrift Für Psychologie / Revue Suisse de Psychologie*, *58*(4), 233–240. https://doi.org/10.1024/1421-0185.58.4.233

Rothschild, J. E., Howat, A. J., Shafranek, R. M., & Busby, E. C. (2019). Pigeonholing partisans: Stereotypes of party supporters and partisan polarization. *Political Behavior*, *41*(2), 423–443. https://doi.org/10.1007/s11109-018-9457-5

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., Bursalıoğlu, A., Bühler, V., Čadek, M., Çetinçelik, M., Clay, G., Cortijos-Bernabeu, A., Damnjanović, K., Dugue, T. M., Esberg, M., Esteban-Serna, C., … Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour*, *5*(10), 1369–1380. https://doi.org/10.1038/s41562-021-01092-x

Rydell, R. J., Hugenberg, K., Ray, D., & Mackie, D. M. (2007). Implicit theories about groups and stereotyping: The role of group entitativity. *Personality and Social Psychology Bulletin*, *33*(4), 549–558. https://doi.org/10.1177/0146167206296956

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.1910.01108

Shehata, A. (2014). Game frames, issue frames, and mobilization: Disentangling the effects of frame exposure and motivated news attention on political cynicism and engagement. *International Journal of Public Opinion Research*, *26*(2), 157–177. https://doi.org/10.1093/ijpor/edt034

Simas, E. N., Clifford, S., & Kirkland, J. H. (2020). How empathic concern fuels political polarization. *American Political Science Review*, *114*(1), 258–269. https://doi.org/10.1017/S0003055419000534

Sjöström, A., & Gollwitzer, M. (2015). Displaced revenge: Can revenge taste "sweet" if it aims at a different target? *Journal of Experimental Social Psychology*, *56*, 191–202. https://doi.org/10.1016/j.jesp.2014.09.016

Spencer-Rodgers, J., Hamilton, D. L., & Sherman, S. J. (2007). The central role of entitativity in stereotypes of social categories and task groups. *Journal of Personality and Social Psychology*, *92*(3), 369–388. https://doi.org/10.1037/0022-3514.92.3.369

Statista. (2025). *U.S. top social media sites visit share 2025*. https://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/

Suhay, E., Brandt, M. J., & Proulx, T. (2017). Lay belief in biopolitics and political prejudice. *Social Psychological and Personality Science*, *8*(2), 173–182. https://doi.org/10.1177/1948550616667615

Torcal, M., & Thomson, Z. A. (2023). Social trust and affective polarization in Spain (2014–19). *Electoral Studies*, *81*, Article 102582. https://doi.org/10.1016/j.electstud.2023.102582

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, *69*, Article 102199. https://doi.org/10.1016/j.electstud.2020.102199

Walton, G. M., & Banaji, M. R. (2004). Being what you say: The effect of essentialist linguistic labels on preferences. *Social Cognition*, *22*(2), 193–213. https://doi.org/10.1521/soco.22.2.193.35463

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., & Poli, I. (2024). *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*. arXiv. https://doi.org/10.48550/arXiv.2412.13663

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yzerbyt, V. Y., Rogier, A., & Fiske, S. T. (1998). Group entitativity and social attribution: On translating situational constraints into stereotypes. *Personality and Social Psychology Bulletin*, *24*(10), 1089–1103. https://doi.org/10.1177/01461672982410006

CHAPTER 5
General Discussion

CHAPTER 5

In the General Introduction I have argued that previous research has thus far largely overlooked social categorization as a crucial prerequisite for many of the presumed negative consequences of affective polarization. The negative stereotypes and prejudice attached to political outgroups translate into lived experience and behavior only to the extent that individuals perceive each other as members of political outgroups. Yet individuals do not automatically know which category someone belongs to. Unless a person explicitly states their political affiliation, political categories can only be inferred, for instance from opinions, behavior, or other cues. But do people make these inferences in everyday life? The results presented in Chapter 2 suggest that they do: Participants spontaneously inferred a wide variety of political categories from relatively subtle cues. As we argued, these inferences likely reflect high perceived diagnosticity, with cues being used for categorization when they are perceived as uniquely tied to one political group. In Chapter 3 we showed that perceived diagnosticity is malleable: Through counterstereotypical exemplars, cues can be made to appear less diagnostic of category membership, reducing their use in categorizing others. In Chapter 4 we provided evidence for the malleability of a further aspect of group perception, namely the degree to which political groups are perceived as entitative. Our analysis of social media posts suggests that the use of linguistic labels can amplify the effects of affective polarization by making categories appear as entitative groups rather than loose collections of individuals. Taken together, the three articles suggest that citizens often perceive each other as members of entitative political groups—that is, as partisans rather than individuals. Crucially, however, they also suggest that this perception is malleable.

Because each article has already addressed the implications and future directions specific to its core construct, this general discussion focuses on the insights that emerge from considering them together. I first examine how spontaneous categorization, diagnosticity, and entitativity interact, and what further research questions arise from taking this joint perspective. I then discuss how these constructs fit into broader polarization dynamics and whether they constitute a driver of polarization

or a mere epiphenomenon. Finally, I consider to what extent they may serve as leverage points for attempts to reduce polarization.

Interdependencies Between Core Constructs

While each construct has been examined individually in the preceding articles, a closer examination reveals meaningful conceptual overlap between them. Understanding these interdependencies not only clarifies the theoretical landscape and fruitful directions for future research but also points toward more effective avenues for intervention efforts. In what follows, I examine three key relationships: the interplay between spontaneous categorization and diagnosticity, the connection between entitativity and diagnosticity, and finally, the relationship between spontaneous categorization and entitativity.

Interplay of Spontaneous Categorization and Diagnosticity

In our first article, we examined categorization processes and argued that their robustness may stem from the fact that in polarized contexts many cues are—or at least appear to be—highly diagnostic of political categories. A key methodological strength was our use of indirect impression formation measures, which demonstrated that political categorization occurs spontaneously without the explicit prompts present in direct measures (like rating scales). This approach creates processing conditions that should more closely resemble everyday life, where we typically observe others without being asked to evaluate whether specific categories apply to them. The choice between direct and indirect measures matters, because they can elicit different cognitive processing patterns (Ma et al., 2011), and in some domains have produced clear dissociations (e.g., Van Dessel et al., 2021). Our second article revisited and empirically investigated the idea that diagnosticity influences categorization but assessed categorization only using direct measures, thereby creating a gap between the two articles. We chose direct measures for a practical reason: Standard impression formation paradigms (Carlston & Skowronski, 1994; Todorov & Uleman, 2002; Uleman et al., 1992) typically require many different cues, considerably more than we could realistically manipulate for

diagnosticity using our method. Nevertheless, this practical constraint means we have yet to test whether diagnosticity also guides spontaneous impression formation.

It seems plausible that diagnosticity would also guide spontaneous categorization, as this is implied in several rational stereotype/categorization models (e.g., Gershman & Cikara, 2023; Solanki & Cesario, 2025). Moreover, findings by Mitchell et al. (2006) suggest that spontaneous categorization may be especially dependent on diagnosticity, because only diagnostic information was found to trigger socio-cognitive processing in the absence of explicit impression formation instructions. However, Mitchell et al. (2006)—like many others (Cone & Ferguson, 2015; Geeraert & Yzerbyt, 2007; Ginosar & Trope, 1980; Krueger & Rothbart, 1988; Rasinski et al., 1985)—manipulated diagnosticity between cues, leaving potential stimulus confounds unaddressed. This limitation is significant because some scholars argue that categorization often operates pseudodiagnostically (Skowronski, 2002). Drawing on an analogy to scientific hypothesis testing, pseudodiagnostic hypothesis testing is understood as a strategy in which we focus on evidence for our hypothesis while neglecting evidence for alternative hypotheses (Trope & Liberman, 1996). In our terminology from Chapter 3, this corresponds to focusing on the stereotype probability while neglecting the counterstereotype probability, or categorizing based on prototypicality/representativeness rather than diagnosticity. Some researchers argue that social hypothesis testing is typically pseudodiagnostic unless observers are prompted to weigh an actor's membership across different categories through contradictory information (Skowronski, 2002). For example, when asked how likely a target with a strong anti-abortion stance is to be a Republican, observers would test pseudodiagnostically, considering only how well this stance fits their stereotype of Republicans. Conversely, when the target has one stereotypically Republican (e.g., anti-abortion) and one stereotypically Democratic (e.g., pro-gun-control) stance, observers would weigh which of the two contradictory stances is more diagnostic of partisanship.

Given that diagnostic testing should be more computationally intensive than pseudodiagnostic testing (Trope & Liberman, 1996), and that neuroimaging studies generally report

lower activity for spontaneous than for deliberative categorization (Ma et al., 2011; Mitchell et al., 2006), one might speculate that spontaneous categorization is characterized by pseudodiagnostic testing and therefore guided by prototypicality rather than diagnosticity. However, this line of reasoning relies on the assumption that category hypotheses are already available, and spontaneous categorization is merely about hypothesis testing. By contrast, spontaneous categorization is characterized precisely by the absence of such hypotheses. In the absence of explicit instructions to judge the likelihood that a person belongs to a particular category, hypotheses must first be generated rather than merely tested. Consequently, even if hypothesis testing were pseudodiagnostic, diagnosticity could still shape spontaneous categorization at the earlier stage of hypothesis generation. Considering these constructs jointly thus points to an important avenue for future research. Specifically, studies should investigate whether and under what circumstances spontaneous categorization is guided by these two variables, following our approach of separately manipulating diagnosticity and prototypicality within the same behavior to exclude stimulus confounds.

It is important to note that the distinction between diagnosticity and prototypicality is likely not equally significant across all contexts. Our mental representations of categories primarily contain information that is diagnostic for at least one relevant contrast (Ford & Stangor, 1992; Solanki & Cesario, 2025). This means that attributes which are common for a category but also apply to all other categories become associated with the categories (i.e., become stereotypes). In other words, the attributes that become stereotypes are primarily those that are diagnostic for at least one categorical contrast. In contexts where there is only one relevant contrast, such as the American two-party landscape (Republicans vs. Democrats), the correlation between diagnosticity and stereotypicality should therefore always be relatively high: An attribute that is a stereotype for one party is unlikely to also be a stereotype for the other party. However, the more complex the contrast structure of the category system becomes (for example, in multiparty systems) and the more partially overlapping categories exist, the more likely stereotypes will emerge that are highly

prototypical of certain categories but only diagnostic of a particular contrast rather than any specific category. For instance, in the German context strong support for traditional family values may not be diagnostic of any singular party, but for the contrast between AfD/CDU/CSU versus Grüne/Linke/SPD. The question of whether spontaneous categorization is guided by diagnosticity or prototypicality is thus likely more relevant for multiparty systems.

Considering the interplay of spontaneous categorization and diagnosticity jointly in the multi-party context therefore gives rise to another important avenue for future research: If spontaneous categorization is indeed guided by diagnosticity, future research should assess whether this leads to reduced party categorization in multiparty systems, perhaps in favor of more simplistic ideological (left-right) categorization. Alternatively, parties may still be inferred even in multiparty systems, but when two parties are equally associated with a given cue, other factors—such as category accessibility (Newman & Uleman, 1990)—may guide the selection process (for a discussion of accessibility, see the section on the interplay of entitativity and spontaneous categorization below).

Interplay of Diagnosticity and Labeling/Entitativity

Diagnosticity is also related to entitativity, albeit indirectly. Perceptions of entitativity are thought to be driven by a multitude of different factors. Among the most important antecedents discussed in the literature are perceptions of similarity between group members, interaction between group members, common goals, and common outcomes (Blanchard et al., 2020; Campbell, 1958; Gaertner et al., 2006; Lickel et al., 2000). Crucially, not all of these antecedents have to be present for perceptions of entitativity to arise. Rather, many authors suggest that groups can acquire their entitativity through various pathways, and that the antecedents correlate with entitativity, but can nevertheless be separated from it (Blanchard et al., 2020; Campbell, 1958; Kashima et al., 2005; Lickel et al., 2000; Spencer-Rodgers et al., 2007). Consistent with this perspective, evidence indicates that different antecedents of entitativity can sometimes be manipulated independently (e.g., Crump et al., 2010; but see Dang & Liu, 2022) and can have distinct effects on group perception (Dang & Liu,

2022; Insko et al., 2013). For example, Dang and Liu (2022) found that a manipulation of increased similarity led to polarization of competence evaluations of a group (competent groups becoming more competent and incompetent groups becoming more incompetent), whereas a manipulation of interdependence (interaction, common goals, common outcomes) led to a general increase in competence judgments.

When we consider the antecedents of entitativity separately—as this research suggests we should—a clear connection to diagnosticity emerges. Specifically, group similarity is logically linked to diagnosticity: The more category members adopt a given cue, the more similar they become with respect to it, and thus the more diagnostic that cue becomes of the category. This is illustrated by findings from Novoa et al. (2023, Study 3), who presented participants with generic statements about novel but plausible-sounding preferences of Democrats and Republicans (e.g., "Democrats support House Bill 858", p. 8) and then asked them about the perceived prevalence of these preferences. They found that generic statements led to an estimated prevalence of almost 70%, about 4% more than non-generic statements referring to "many Republicans" and 31% more than ones referring to "some Republicans". These heightened prevalence estimates resulting from generic compared to non-generic expressions imply both greater similarity between group members (i.e., lower group variance) with regard to the preference as well as higher diagnosticity of the preference for the respective category. Thus, considering diagnosticity and entitativity jointly shows that perceived group similarity is doubly relevant. Perceiving low group variance should not only drive perceptions of entitativity but also make group stereotypes appear more diagnostic.

Given that Americans strongly underestimate the variance of partisans' policy stances (Dias et al., 2025), it is paramount to investigate interventions aimed at correcting this misperception or, more generally, at increasing perceived variance. There is, for instance, growing evidence that this can be achieved through specific types of data visualizations, with positive effects on polarization-related outcomes (Markant et al., 2026; Wolf & Hanel, 2024). For example, Markant et al. (2026) found that visualizations correcting outparty perceptions by displaying the full distribution of

attitudes rather than just the mean were more effective in reducing Americans' support for political violence and undemocratic practices. With regard to linguistic labeling as a driver of group variance perceptions, the findings by Novoa et al. (2023) suggest that generics may be particularly potent, as they not only increase the perceived prevalence of ascribed attributes for the mentioned group but also decrease the perceived prevalence for the unmentioned group. Generic statements thus signal not only that the labeled group is similar with respect to an attribute, but also that the unmentioned group is similar in lacking it. In other words, generics may simultaneously enhance the perceived entitativity of both groups as well as the diagnosticity of the attribute. Interventions aimed at reducing the use of generics may therefore be particularly effective in reducing both the likelihood of categorization and the severity of its consequences. However, it remains to be seen how these findings translate to contexts with more than two groups: In multi-party systems the assumption that generically ascribed attributes are distinctive to the mentioned group may be less pronounced, as people may expect generic statements to signal diagnosticity of an unspecified categorical contrast rather a particular party. In this case, the effect of generics on perceived prevalence may be limited to the mentioned group.

Beyond reducing generic language, it may also be fruitful to investigate more explicitly corrective interventions. Though the comparison may be unwelcome on both sides of the aisle, research on child-directed speech offers some intriguing suggestions. Foster-Hanson et al. (2022) presented children with generic statements about fictional categories (e.g., "Zarpies have striped hair", p. 23), which were subsequently corrected by a more knowledgeable speaker who either rejected the claim as false of the category in general ("upholding the scope"; e.g., "No, that's not true about Zarpies", p. 23) or restricted it to the individual instance ("limiting the scope"; e.g., "No, *this* Zarpie has striped hair", p. 23). The latter strategy led to reduced subsequent use of the category, possibly because it individualized the target rather than merely denying the generalization. This approach may translate well to adult political discourse because it helps maintain the flow of the conversation. Rather than rejecting a generic claim outright (e.g., "I don't think Democrats are

hypocritical") which could be perceived as confrontational, a scope-limiting response (e.g., "Well, this Democrat is certainly hypocritical") maintains the focus on the individual case that prompted the generalization while resisting its extension to the broader category. Future research could therefore compare different corrective strategies for responding to generic political generalizations in everyday contexts.

Interplay of Entitativity and Spontaneous Categorization

Finally, entitativity may be indirectly related to spontaneous categorization—via linguistic labeling as a shared antecedent. Beyond its effect on entitativity, linguistic labeling should also affect the accessibility of social categories. Accessibility is defined as "the activation potential of available knowledge" (Higgins, 1996, p. 134). A highly accessible category thus has a higher potential to be activated when an individual is presented with a relevant cue, such as a diagnostic behavior. Accessibility has been shown to be driven by recent and frequent activation, for instance through the brief presentation of category labels (Higgins, 1996). Findings by Newman and Uleman (1990, Study 1) illustrate this effect: They presented participants with descriptions of people (e.g., "Molly would not take no for an answer", p. 230), that had been shown to be equally diagnostic of an evaluatively positive (persistent) and an evaluatively negative (stubborn) trait, and then assessed which of the two traits was spontaneously inferred from these descriptions. Importantly, participants had been primed with one of the two trait words beforehand and results showed that (at least when participants were unaware of the prime) they more frequently inferred the primed trait. In the domain of political communication, the use of political group labels may likewise function as a form of priming and lead to increased accessibility of the respective categorical representations. Someone who has spent half an hour in the morning reading newspaper articles about the latest misdeeds of liberals or conservatives may therefore be more likely to infer these categories throughout the day when encountering more or less diagnostic categorization cues.

The fact that repeated stimulation of categorical representations may increase their accessibility even further, potentially even making them chronically elevated (Higgins, 1996), once

again points to the important role of public media coverage with its frequent use of group labels and partisan framing. Non-partisan or issue-focused framing that refrains from using political labels altogether may help achieve lower levels of situational or perhaps even chronic category accessibility and thereby could lead to reduced political categorization. Moreover, compared to so-called game-frame coverage emphasizing strategic political motivations, issue-focused framing has been found to reduce political cynicism and increase political interest (Shehata, 2014) and may additionally lead to even lower perceptions of entitativity than non-generic adjective labels (see Foroni & Rothbart, 2013).

Newer frameworks suggest that apart from accessibility, motivational relevance may also have an important effect on category activation (Eitam & Higgins, 2010; Higgins & Eitam, 2014). For instance, the extent to which "a representation is perceived as real (in an epistemic sense)" is argued to affect how easily it is activated (Higgins & Eitam, 2014, p. 233). This suggests that entitativity might also have a direct effect on spontaneous categorization. Although research on spontaneous trait inferences (STIs) suggests that high entitativity is probably not a necessary precondition for spontaneous categorization, seeing as trait categories should have relatively low entitativity and are nevertheless reliably inferred from diagnostic behaviors (Bott et al., 2024; Todorov & Uleman, 2002), entitativity may nevertheless amplify categorization tendencies. To my knowledge, no studies have investigated this so far, making this another potential avenue for future research.

Driver or Epiphenomenon

Having examined the interdependencies between the core constructs, I now turn to how political categorization and perceived entitativity, which I will collectively refer to as group perception, fit into the broader polarization process. To this end, I discuss the potential causal effects of group perception on the other aspects of polarization introduced in Chapter 1. I thereby probe the question whether group perception should be viewed as a mere epiphenomenon of polarization, or whether it can instead be considered one of its drivers.

Before turning to the specific relationships between group perception and other aspects of polarization, however, I first address a more general question: Under what circumstances are categorization and its effects likely to occur? This matters because the extent to which individual-level effects can account for changes in societal-level phenomena depends on how frequently those effects are evoked. If, for instance, categorizing someone as a political outgroup member were to produce more extreme political attitudes, this would only translate into a meaningful societal-level effect (i.e., a causal effect of the inferability of political categories on ideological polarization) if the underlying pathway were frequently realized in everyday life. It is therefore essential to identify the situational characteristics that make categorization-based effects more or less likely, as this allows us to gauge how frequently individual-level effects are realized.

The social cognition literature on person perception helps specify the situations in which political categorization is likely to be consequential. The Continuum Model (Fiske, 2018; Fiske & Neuberg, 1990) posits that although perceivers may initially categorize others upon encountering them, categorization only results in a category-based construal if no category-incongruent information is available. When the available information does not fit neatly into a categorical representation, the target is instead individuated—that is, their different attributes, including category membership, are considered in an integrated fashion. Consistent with this, studies have shown that providing participants with individuating information about target people reduces their reliance on social categories in subsequent evaluations (Navon et al., 2026; Rubinstein et al., 2018). In line with the idea that people store a lot of individuating information about familiar targets, Quinn et al. (2009) found that presentation of familiar compared to unfamiliar target faces led to comparably less gender-based categorization. Interestingly, this was already observable in the reduced automatic *activation* of gender categories (see also Habibi & Khurana, 2012). Taken together, these findings suggest that categorical construal is most likely with unfamiliar others about whom little to no individuating information is available. This is not to say that categorical construal is impossible when more information is available—in principle, all knowledge about a person could be

consistent with a categorical representation. But the more information we have about a person, the higher are the chances that some of it will cut against the categorical grain.

A second situational factor concerns whether whatever individuating information does exist is available independently of the perceiver's own behavior. Allidina et al. (2025) argued that in many contexts, individuating information is only accessible if perceivers actively choose to approach the target. They found that in such contexts, perceivers individuated outgroup members less than in contexts where individuating information was available regardless of perceiver behavior (at least when interactions with outgroup members tended to be negative). In other words, contexts where individuating information only becomes available through own effort or potentially aversive interactions (e.g., reading a stranger's social media comment endorsing an outparty viewpoint) are more conducive to categorical construal than contexts where individuating information is readily available from the outset (e.g., reading a feature story about an outparty politician). Taken together, environments are most conducive to categorization and its effects when they involve frequent encounters with unfamiliar others, are rich in diagnostic cues, offer little individuating information, and present whatever individuating information does exist only after categorical information and contingent on approach. With this in mind, the following sections examine how group perception may causally relate to other aspects of polarization.

Group Perception and Ideological Polarization

In their widely cited definition, McCoy et al. (2018) described polarization as "a process whereby the normal multiplicity of differences in a society increasingly align along a single dimension, cross-cutting differences become instead reinforcing, and people increasingly perceive and describe politics and society in terms of 'Us' versus 'Them'" (p. 18). This definition implicitly frames group perception as a consequence of ideological polarization. This view makes intuitive sense as perceiving a reduced overlap between political camps provides a rational basis for group perception—although, as I have argued, citizens routinely fail to perceive this overlap accurately (Ahler, 2014; Ahler & Sood, 2018; Dias et al., 2025; Druckman et al., 2022; Lees & Cikara, 2020;

Levendusky & Malhotra, 2016; Moore-Berg et al., 2020; Ruggeri et al., 2021; Westfall et al., 2015; Yudkin et al., 2019). But even if ideological divergence and alignment are overestimated, it nevertheless seems highly plausible, that these exaggerated perceptions are to some extent grounded in actual distributions of beliefs.

There is, however, a lot of research suggesting that the reverse causality may also be true, with group perception affecting the processes via which citizens form political opinions. This starts with what people pay attention to. For example, a study found that when given the choice between two political speeches, higher salience of the speaker's partisan identity leads to a stronger preference for the inparty speaker (Arendt et al., 2023). It also extends to attitude formation and persuasion processes (Mackie & Queller, 2000), with studies showing that when policies are presented with cues about whether one's in- or out-party supports the policy, partisans' own policy attitudes shift away from their out-party or toward their in-party (Bakker et al., 2020; Clifford et al., 2024). Reinero et al. (2023) found similar effects when investigating misinformation, with partisans believing misinformation from ingroup members to a much higher degree than misinformation from outgroup members and misinformation corrections more often backfiring when coming from outgroup compared to ingroup sources. Taken together, the literature suggests that through affecting attitude formation and the likelihood of persuasion, group perception can contribute to ideological polarization, making people attend to, adopt, or be persuaded by ingroup opinions to a greater extent than outgroup opinions.

While the research reported in the previous section relied on rather blatant party cues, we demonstrated in Chapter 2 that political categories can also be spontaneously inferred from much subtler signals, including expressed political opinions. This creates a potential challenge for attitude change when the attitudes that individuals wish to promote are themselves diagnostic of group membership. For example, if person A attempts to persuade person B of the importance of combating climate change, B may infer that A belongs to an outgroup (e.g., liberals) and consequently discount A's arguments. It seems plausible that even the use of specific words may in

some cases be perceived as diagnostic of group membership and thereby shape opinion formation. A recent anecdote from German public discourse illustrates this dynamic. The term “remigration” (“Remigration”) recently became a political buzzword after an investigative report revealed the radicalized form in which participants at a private meeting of far-right figures had discussed the concept, extending it beyond voluntary return to the potential forced removal of individuals considered culturally or ethnically undesirable. As a result, the previously neutral term became widely associated with the far right and may now function as a diagnostic cue of far-right affiliation. To the extent that the use of certain words becomes diagnostic of political group membership, our inability to find common ground might set in even before we share our opinions—and possibly irrespective of them. Future research should examine whether such indirect cues of group membership influence phenomena such as news selection and persuasion in ways similar to the more explicit cues commonly studied.

An interesting arena for this endeavor would be the study of moral reframing (Feinberg & Willer, 2013, 2015; Gelfand et al., 2022; Jansson & Strimling, 2025). The literature on moral reframing shows that liberals and conservatives prioritize different moral values and that messages are more persuasive when they use language that reflects the ones held by the target audience. For instance, Feinberg and Willer (2015) found that conservatives were more strongly persuaded to support same-sex marriage by arguments appealing to the moral value of loyalty (i.e., emphasizing that gay couples are patriotic Americans) than by arguments appealing to fairness (i.e., emphasizing that all citizens deserve equal treatment and that denying same-sex marriage constitutes direct discrimination). The authors attributed these effects to a greater moral fit between message and recipient. My present reasoning suggests an alternative mechanism that to date remains untested: Rather than increasing persuasion solely by aligning messages with recipients’ core moral beliefs, morally reframed messages may also or even predominantly signal shared group membership. As argued above, such signaling could itself contribute to the persuasive advantage of morally congruent messaging.

While the extent to which such subtle cues to group membership influence persuasion is still an open question, the general effect of group membership—and partisanship in particular—on belief or attitude formation and persuasion is well established. By pulling individuals toward ingroup opinions and pushing them away from outgroup stances, perceiving others as group members rather than individuals can affect what people believe. As we have argued, if this individual-level effect is frequently evoked in everyday life, this would support the idea that also at the societal level, group perception (or more specifically, high inferability of political categories) may have a causal effect on ideological polarization. Based on my reasoning from the previous section, a frequent realization of this effect appears highly plausible. A lot of political communication occurs via traditional or social media, where citizens digest opinionated political commentaries, watch debates featuring unfamiliar candidates, read about strangers' views in online comment sections, or encounter political messages in campaign advertising. These contexts are characterized by unfamiliar communicators about whom little individuating information is available (and if so, usually contingent on approach), and salient diagnostic or even explicit ideological or party cues, thus fulfilling the conditions conducive to categorical person perception. This suggests that at the societal level, group perception is not merely a byproduct of ideological alignment but a likely mechanism through which opinions and identities become increasingly aligned and perhaps more extreme over time.

Group Perception and Affective Polarization

Now I turn to the societal-level relationship between group perception and affective polarization. As described above, affective polarization drives social avoidance across a number of different domains including online dating, hiring decisions, and roommate or neighborhood selection (Gift & Gift, 2015; Gimpel et al., 2025; Huber & Malhotra, 2017; Inbar & Lammers, 2012; Shafranek, 2021) and may therefore affect the amount of intergroup contact. This is important because positive intergroup contact is an important remedy against prejudice, the effects of which have been demonstrated across different group domains (Paluck et al., 2019), including partisanship (Levendusky & Stecula, 2023; Wojcieszak & Warner, 2020). Affective polarization may reduce the

amount of such positive contact and thereby sustain itself over time. Of course, not all intergroup contact is positive. Allport (1954) suggested that positive contact is more likely given equal status between groups, a common goal, between-group cooperation, and institutional support. MacInnis and Page-Gould (2015) suggested that contact may often be experienced negatively initially and become more positive through longer and repeated interactions. In line with this, enduring forms of contact, such as friendships, have been found to be among the best predictors of reduced prejudice (Pettigrew & Tropp, 2006). In the political domain, Levendusky and Stecula (2023) additionally proposed civility as an important precondition for positive contact—that is, the free and respectful exchange of ideas, with individuals genuinely listening rather than dismissing opposing views. In the absence of these preconditions, contact can often be negative and lead to increases in prejudice and related outcomes (Aberson, 2015; Allport, 1954; Árnadóttir et al., 2018; Barlow et al., 2012; Hayward et al., 2017). By reducing empathy and trust toward outpartisans (Carlin & Love, 2013; Cikara et al., 2011; Robbett et al., 2023; Simas et al., 2020), affective polarization may, beyond reducing contact, also make the remaining contact more negative. This suggests that affective polarization may not just be self-sustaining but rather self-reinforcing.

Crucially, for affective polarization to affect the quality or quantity of contact, citizens need to construe each other as outgroup members. Categorization enables individuals to turn their dormant partisan bias into manifest social avoidance behaviors or negative contact experiences. Whether group membership is inferred from opinions, group identity symbols, or explicit party labels—whenever it is in hand, it can drive individuals' decisions to self-select out of intergroup contact or impoverish its quality. I therefore argue that at the societal level, rather than driving affective polarization directly, group perception determines how well affective polarization can sustain itself. A similar dynamic is theorized by Törnberg et al. (2021), who model polarization as a process resulting from an imbalance between fragmenting and integrating forces. They argue that individuals have a tendency to self-segregate into homogeneous groups that has historically been counterbalanced by institutions that confront us with dissimilar others and facilitate positive

outgroup contact (e.g., schools, workplaces, neighborhoods, or associations). According to their model, tipping this balance in favor of the fragmenting forces can produce a self-reinforcing polarization process. My framework elaborates on this idea by positing that heightened group perception may be a factor that tips this balance toward fragmentation.

As I have argued earlier, individual-level effects need to be frequently evoked in everyday life to support the idea of corresponding societal-level effects. That is, group perception can only meaningfully increase the degree to which affective polarization is self-reinforcing if there are sufficient situations in which categorization can prevent positive contact. I argued that social media may offer plenty of situations conducive to categorization-based effects, given that it is rich in diagnostic cues, and low in individuating information. But while social media may facilitate negative intergroup interaction—for instance through uncivil interactions in comment sections (see Wu & Resnick, 2021)—it is not a context where people are most likely to form close relationships, which usually develop in contexts with a high likelihood of repeated, unplanned, face-to-face interaction (Reagans, 2011; Small & Adler, 2019) such as in education, at the workplace, in neighborhood settings, or in voluntary or religious organizations (Thomas, 2019). These contexts may differ with regard to how conducive they are to group-based person construal. For instance, in sports associations or workplaces requiring strong collaboration in small teams, it is harder to opt out of contact even when individuals are categorized at a very early stage, which enables individuation at a later stage. In neighborhood contexts, more loosely structured work environments, or higher education, on the other hand, there is much greater flexibility in choosing interaction partners. Categorization cues may be present from early on and in the absence of external factors enforcing contact may lead to social avoidance. While this would suggest that schools, sports associations, (some) workplaces, voluntary organizations, and so on, constitute an integrating force by enabling positive intergroup contact, this is not necessarily true, as already the selection of these contexts can be driven by categorization of the organizations as these increasingly signal party support via campaign contributions or by taking stands on salient political issues (Roth et al., 2022). Thus,

whether through enabling the avoidance of cross-cutting contexts or of cross-cutting interactions within contexts, categorization effects may plausibly prevent positive intergroup contact across a wide range of everyday situations. Therefore, group perception is likely to not just be an epiphenomenon of polarization but to have a causal influence via increasing the degree to which affective polarization is self-reinforcing.

Within this section I thus far focused broadly on the role of group perception. There is, however, an additional mechanism through which entitativity in particular may affect affective polarization. We already briefly described these mechanisms in Chapter 4, but it is worth reiterating them here: First, high entitativity has been shown to have a causal effect on partisan prejudice (Hoyt et al., 2026), a link that is supported by the results of our social media analysis from Chapter 4. Second, although we could not confirm this effect in our analyses, previous research has shown entitativity to strengthen individual-to-group generalization (Crawford et al., 2002; see also Effron et al., 2018; Pereira & Van Prooijen, 2018). Given that extreme group members are often overrepresented in public discourse (Eady et al., 2025; Padgett et al., 2019), high entitativity could contribute to the emergence of exaggerated stereotypes—a key feature of affective polarization (Ahler, 2014; Druckman et al., 2022; Levendusky & Malhotra, 2016; Westfall et al., 2015; Yudkin et al., 2019)—by increasing the degree to which extreme attributes are generalized to the entire category. A recent controversy described in the New York Times (Bensinger & Thompson, 2025) illustrates this dynamic well: After a small number of critics accused a jeans ad campaign (“Sydney Sweeney has great jeans”) of carrying white supremacist undertones, right-wing commentators and politicians quickly framed this fringe reaction as representative of the left broadly, fueling sustained outrage. Whether this framing reflected genuine generalization from a minority view to the group as a whole, or was strategic, the episode demonstrates how readily unrepresentative voices can be perceived as typical of an entire group. High perceived entitativity may have facilitated these generalizations or increased the degree to which they seemed credible to others. Dynamics like this may play out frequently, especially in online or media contexts conducive to group-based person

construal, constituting a route through which group perception may directly contribute to affective polarization.

To summarize, I have illustrated several pathways through which at the societal level group perception may have a causal effect on the more widely studied aspects of polarization. Specifically, I argued that group perception can increase ideological polarization through shaping attitude formation and persuasion and increase affective polarization through influencing the amount and quality of intergroup contact as well as through individual-to-group generalization. I cited evidence supporting the individual-level effects in which the presumed societal-level relationships are grounded as well as arguments for why a frequent realization of these effects seems plausible. I therefore conclude that group perception should be viewed not as a mere epiphenomenon, but rather as a likely driver of polarization.

Suitability as Leverage Point

After establishing group perception as a likely causal factor in polarization dynamics, the question now becomes whether it also constitutes a promising leverage point—that is, whether it is manipulable enough to be worth targeting for prevention and intervention. Across this dissertation, I addressed various avenues for reducing group perception, focusing mainly on changing individuals' perceptions. For instance, I suggested that the perceived diagnosticity of political opinions can be reduced by motivating individuals to voice their cross-cutting opinions, which on social media platforms may be achievable through stronger community moderation (Matias, 2019). I also put forward that perceived entitativity of political groups can be reduced by changing the way these groups are labeled in public discourse, which may be achievable through the implementation of journalistic guidelines in news media organizations, updated feed-ranking algorithms that down-rank social media posts featuring generic or noun labels (see Jia et al, 2024), or perhaps nudges prompting social users to reconsider the generality of their posts (e.g., "Does this apply to all Republicans?"; see Pennycook et al., 2021; Pennycook & Rand, 2022). Moreover, I highlighted generic labeling as a particularly promising intervention target as it should drive perceptions of both

diagnosticity and entitativity simultaneously. Taken together, this illustrates that there are multiple potential pathways for reducing group perception by targeting perceived cue diagnosticity as well as perceived entitativity. However, even though the research showing that people routinely overestimate the ideological distance between groups or the homogeneity within groups suggests considerable potential for interventions targeting perceptions, the effects of those interventions will to some extent nevertheless be constrained by how entitative groups or diagnostic cues actually are.

A more effective way than changing how individuals interpret cues, should therefore be to remove categorization cues from the environment altogether. Cue removal seems most feasible for more explicit cues: Partisans could pull up their yard signs, peel off their bumper stickers, or take off their MAGA hats, just as political parties could opt for campaigning strategies that rely less heavily on political merchandise. Citizens could remove political slogans from their social media bios and clear identity labels and symbols from their dating profiles, just as platforms could be designed to inhibit the display of these cues. Talk shows could feature non-partisan experts rather than explicitly labeled politicians, and journalists could refrain from describing policies through a partisan frame, enabling arguments to speak for themselves, at least for as-yet-unbranded issues. Where cues serve no function beyond identity signaling, simple removal may be a low-cost and underappreciated intervention target.

Not all cues, however, are amenable to simple removal. Many behaviors that can function as diagnostic cues—such as using "politically incorrect" language, driving an SUV or pickup truck, carrying a firearm, not wearing masks during a pandemic, indicating preferred pronouns in email signatures, ordering plant-based milk, or using gender-inclusive language—are ones people engage in because they see them as serving an important purpose, not merely to signal their identity. Many behaviors may also be instrumental to certain political goals. For instance, driving a cargo bike may often serve to promote pro-environmental values and using gender-inclusive language may serve to achieve greater gender equality. To the extent that one does not want to obstruct these social developments, removing the cues may not be a desirable goal. Instead, one may want to make the

cues spread quickly and evenly, minimizing the time they serve as categorization cues. When political groups adopt new behavioral norms quickly and at similar rates, this makes these norms less diagnostic and shortens the window in which they can elicit categorizations. But when uptake is slow and asymmetric—with one group largely adopting while the other largely does not—that window stretches, and in the worst case the behavior becomes an entrenched and frequent diagnostic categorization cue.

The most powerful lever for achieving a fast and symmetric uptake may be to enforce behaviors through policy. For instance, standardizing the inclusion of preferred pronouns on social media profiles or introducing emissions standards that force people to buy eco-friendly cars can lead to an accelerated and near-universal uptake of the respective cues. While enforcement may often technically raise the diagnosticity of the deviant behavior (assuming that defiance only occurs in the opposing group), it should dramatically reduce its frequency and thereby nevertheless reduce categorization. Policymakers may often refrain from bans or mandates for fear of public opposition. But reactance to such policies usually decreases significantly within months after implementation (Granulo et al., 2025) and many policies that were initially met with strong opposition, such as seat belt mandates or smoking bans, are now widely accepted (e.g., Hürter, 2024). By implementing policies that enforce certain behaviors as compared to policies aimed at facilitating voluntary behavior change, policymakers may significantly shorten the time period in which the respective behaviors can act as potent categorization cues and thereby prevent their contribution to polarization dynamics. When bans or mandates are infeasible, similar effects may sometimes be achieved when behaviors are very strongly incentivized.

The German heating legislation of 2023 offers a case in point. The legislation required newly installed heating systems to draw at least 65% of their energy from renewable sources, making heat pumps the sensible default technology even without legally mandating them outright. The policy met with intense public backlash, particularly directed at the Green party minister who championed it, yet opposition has since begun to subside and adoption has spread relatively independently of

partisanship (Hajek, 2026). Crucially, because the legislation introduced strong situational demands that made heat pump installation rational even for households with no environmental sympathies, the technology is unlikely to function as a reliable partisan cue in the future. A stricter law prescribing heat pumps as the sole permitted option may have reduced diagnosticity further still. Conversely, in the absence of such legislation, a gradual and self-selected adoption of heat pumps driven by ideological rather than economic considerations could have allowed the technology to become an increasingly frequent, highly visible, and diagnostic neighborhood-level cue of political identity. Whether intentional or not, policy decisions thus shape the ecology of partisan cues that citizens encounter in everyday life—and with it, the ease with which they sort their neighbors, colleagues, and strangers into political camps.

Not every behavior, however, lends itself to institutional enforcement. The use of gender-fair language in everyday speech, for instance, can be neither mandated nor banned. Other behaviors (wearing a face-mask during a pandemic) may be regulable in principle but too difficult to enforce in practice and defiance could thus remain prevalent while potentially being perceived as even more diagnostic as a result. In these cases, political groups, organizations, and individuals face a genuinely difficult strategic choice: whether and when to push for new behavioral norms, weighing the polarizing costs of asymmetric uptake against the substantive benefits of norm change. Where rapid and widespread adoption is unlikely, it may sometimes be wiser to defer norm advocacy to less polarized times—accepting a temporary pause as the price of a less fragmented political environment.

Taken together, these considerations suggest that group perception is not just a likely causal factor in polarization dynamics but also one that can be manipulated in various ways. Through preventing or correcting exaggerated perceptions of group homogeneity, simply removing partisan cues, enforcing a fast and symmetric uptake of new behavioral norms through policy, and similar means, individuals, institutions, and legislators alike can contribute to an information environment in which group perception and its polarizing effects are less likely to occur and citizens move from

seeing each other as partisans to seeing each other as individuals. I therefore conclude that group perception constitutes a highly promising leverage point for interventions aimed at reducing political polarization.

Conclusion

This dissertation has argued that group perception plays an important, yet largely overlooked, role in political polarization. People's tendency to categorize others on the basis of their opinions or behaviors and to view the categorized individuals as a unified group is not a mere byproduct of polarization but a likely driver of it, shaping both ideological and affective polarization through distinct causal pathways. Importantly, I have shown that group perception is malleable: Categorization can be disrupted by counterstereotypical information that reduces the perceived diagnosticity of partisan cues, and the tendency to view categorized outpartisans as a unified collective depends on how they are labeled in political discourse. Beyond these experimentally demonstrated effects, I have argued that the ecology of cues that makes categorization possible can often be powerfully shaped through policy. Group perception thus constitutes a highly promising leverage point for efforts aimed at reducing political polarization.

References

Aberson, C. L. (2015). Positive intergroup contact, negative intergroup contact, and threat as predictors of cognitive and affective dimensions of prejudice. *Group Processes & Intergroup Relations*, *18*(6), 743–760. https://doi.org/10.1177/1368430214556699

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *Journal of Politics*, *76*(3), 607–620. https://doi.org/10.1017/S0022381614000085

Ahler, D. J., & Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, *80*(3), 964–981. https://doi.org/10.1086/697253

Allidina, S., Mack, M. L., & Cunningham, W. A. (2025). Experience shapes the granularity of social perception: Computational insights into individual and group-based representations. *Journal of Experimental Psychology: General*, *154*(8), 2257–2271. https://doi.org/10.1037/xge0001770

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Arendt, F., Northup, T., Forrai, M., & Scheufele, D. (2023). Why we stopped listening to the other side: How partisan cues in news coverage undermine the deliberative foundations of democracy. *Journal of Communication*, *73*(5), 413–426. https://doi.org/10.1093/joc/jqad007

Árnadóttir, K., Lolliot, S., Brown, R., & Hewstone, M. (2018). Positive and negative intergroup contact: Interaction not asymmetry. *European Journal of Social Psychology*, *48*(6), 784–800. https://doi.org/10.1002/ejsp.2365

Bakker, B. N., Lelkes, Y., & Malka, A. (2020). Understanding partisan cue receptivity: Tests of predictions from the bounded rationality and expressive utility perspectives. *The Journal of Politics*, *82*(3), 1061–1077. https://doi.org/10.1086/707616

Barlow, F. K., Paolini, S., Pedersen, A., Hornsey, M. J., Radke, H. R. M., Harwood, J., Rubin, M., & Sibley, C. G. (2012). The contact caveat: Negative contact predicts increased prejudice more

than positive contact predicts reduced prejudice. *Personality and Social Psychology Bulletin*, *38*(12), 1629–1643. https://doi.org/10.1177/0146167212457953

Bensinger, K., & Thompson, S. A. (2025, August 7). How the right shaped the debate over the Sydney Sweeney ads. *The New York Times*. https://www.nytimes.com/2025/08/07/business/sydney-sweeney-ad-right-wing-media.html

Blanchard, A. L., Caudill, L. E., & Walker, L. S. (2020). Developing an entitativity measure and distinguishing it from antecedents and outcomes within online and face-to-face groups. *Group Processes & Intergroup Relations*, *23*(1), 91–108. https://doi.org/10.1177/1368430217743577

Bott, A., Brockmann, L., Denneberg, I., Henken, E., Kuper, N., Kruse, F., & Degner, J. (2024). Spontaneous trait inferences from behavior: A systematic meta-analysis. *Personality and Social Psychology Bulletin*, *50*(1), 78–102. https://doi.org/10.1177/01461672221100336

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*(1), 14–25. (1301269891). https://doi.org/10.1002/bs.3830030103

Carlin, R. E., & Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior*, *35*(1), 43–63. https://doi.org/10.1007/s11109-011-9181-x

Carlston, D. E., & Skowronski, J. J. (1994). Savings in the relearning of trait information as evidence for spontaneous inference generation. *Journal of Personality and Social Psychology*, *66*(5), 840–856. https://doi.org/10.1037/0022-3514.66.5.840

Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, *20*(3), 149–153. https://doi.org/10.1177/0963721411408713

Clifford, S., Leeper, T. J., & Rainey, C. (2024). Generalizing survey experiments using topic sampling: An application to party cues. *Political Behavior*, *46*(2), 1233–1256. https://doi.org/10.1007/s11109-023-09870-1

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology*, *83*(5), 1076–1094. https://doi.org/10.1037/0022-3514.83.5.1076

Crump, S. A., Hamilton, D. L., Sherman, S. J., Lickel, B., & Thakkar, V. (2010). Group entitativity and similarity: Their differing patterns in perceptions of groups. *European Journal of Social Psychology*, *40*(7), 1212–1230. https://doi.org/10.1002/ejsp.716

Dang, J., & Liu, L. (2022). How to be "groupy" matters: Groups with shared traits and shared goals engender distinct patterns of social judgments. *Group Processes & Intergroup Relations*, *25*(1), 44–59. https://doi.org/10.1177/1368430220942294

Dias, N. C., Lelkes, Y., & Pearl, J. (2025). American partisans vastly under-estimate the diversity of other partisans' policy attitudes. *Political Science Research and Methods*, *13*(3), 725–735. https://doi.org/10.1017/psrm.2024.36

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2022). (Mis)estimating affective polarization. *The Journal of Politics*, *84*(2), 1106–1117. https://doi.org/10.1086/715603

Eady, G., Bonneau, R., Tucker, J. A., & Nagler, J. (2025). News sharing on social media: Mapping the ideology of news media, politicians, and the mass public. *Political Analysis*, *33*(2), 73–90. https://doi.org/10.1017/pan.2024.19

Effron, D. A., Kakkar, H., & Knowles, E. D. (2018). Group cohesion benefits individuals who express prejudice, but harms their group. *Journal of Experimental Social Psychology*, *79*, 239–251. https://doi.org/10.1016/j.jesp.2018.08.002

Eitam, B., & Higgins, E. T. (2010). Motivation in mental accessibility: Relevance of a representation (ROAR) as a new framework. *Social and Personality Psychology Compass*, *4*(10), 951–967. https://doi.org/10.1111/j.1751-9004.2010.00309.x

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science*, *24*(1), 56–62. https://doi.org/10.1177/0956797612449177

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, *41*(12), 1665–1681. https://doi.org/10.1177/0146167215607842

Fiske, S. T. (2018). *Social cognition: Selected works of Susan Fiske*. Taylor & Francis Group.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, (23), 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Ford, T. E., & Stangor, C. (1992). *The role of diagnosticity in stereotype formation: Perceiving group means and variances*. *63*(3), 356–367. https://doi.org/10.1037/0022-3514.63.3.356

Foroni, F., & Rothbart, M. (2013). Abandoning a label doesn't make it disappear: The perseverance of labeling effects. *Journal of Experimental Social Psychology*, *49*(1), 126–131. https://doi.org/10.1016/j.jesp.2012.08.002

Foster-Hanson, E., Leslie, S., & Rhodes, M. (2022). Speaking of kinds: How correcting generic statements can shape children's concepts. *Cognitive Science*, *46*(12), Article e13223. https://doi.org/10.1111/cogs.13223

Gaertner, L., Iuzzini, J., Witt, M. G., & Oriña, M. M. (2006). Us without them: Evidence for an intragroup origin of positive in-group regard. *Journal of Personality and Social Psychology*, *90*(3), 426–439. https://doi.org/10.1037/0022-3514.90.3.426

Geeraert, N., & Yzerbyt, V. Y. (2007). Cultural differences in the correction of social inferences: Does the dispositional rebound occur in an interdependent culture? *British Journal of Social Psychology*, *46*(2), 423–435. https://doi.org/10.1348/014466606X162062

Gelfand, M., Li, R., Stamkou, E., Pieper, D., Denison, E., Fernandez, J., Choi, V., Chatman, J., Jackson, J., & Dimant, E. (2022). Persuading republicans and democrats to comply with mask wearing: An intervention tournament. *Journal of Experimental Social Psychology*, *101*, Article 104299. https://doi.org/10.1016/j.jesp.2022.104299

Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review*, *30*(4), 1273–1293. https://doi.org/10.3758/s13423-023-02252-y

Gift, K., & Gift, T. (2015). Does politics influence hiring? Evidence from a randomized experiment. *Political Behavior*, *37*(3), 653–675. https://doi.org/10.1007/s11109-014-9286-0

Gimpel, J. G., Newton, J., & Reeves, A. (2025). Short-haul moves and the political geography of partisanship: Intrametropolitan migration as a force for change in U.S. politics. *Urban Affairs Review*. Advance online publication. https://doi.org/10.1177/10780874251380677

Ginosar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology*, *16*(3), 228–242. https://doi.org/10.1016/0022-1031(80)90066-9

Granulo, A., Fuchs, C., & Böhm, R. (2025). Psychological reactance to system-level policies before and after their implementation. *Proceedings of the National Academy of Sciences*, *122*(18), Article e2409907122. https://doi.org/10.1073/pnas.2409907122

Habibi, R., & Khurana, B. (2012). Spontaneous gender categorization in masking and priming studies: Key for distinguishing Jane from John Doe but not Madonna from Sinatra. *PLoS ONE*, *7*(2), Article e32377. https://doi.org/10.1371/journal.pone.0032377

Hajek, S. (2026, February 19). *Neues Heizungsgesetz: Wärmepumpen sind auch unter AfD-Wählern am beliebtesten*. https://www.wiwo.de/technologie/umwelt/neues-heizungsgesetz-waermepumpen-sind-auch-unter-afd-waehlern-am-beliebtesten/100201339.html

Hayward, L. E., Tropp, L. R., Hornsey, M. J., & Barlow, F. K. (2017). Toward a comprehensive understanding of intergroup contact: Descriptions and mediators of positive and negative contact among majority and minority groups. *Personality and Social Psychology Bulletin*, *43*(3), 347–364. https://doi.org/10.1177/0146167216685291

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). The Guilford Press.

Higgins, E. T., & Eitam, B. (2014). Priming…shmiming: It's about knowing *when* and *why* stimulated memory representations become active. *Social Cognition*, *32*(Supplement), 225–242. https://doi.org/10.1521/soco.2014.32.supp.225

Hoyt, C. L., Burnette, J. L., & Moore, M. (2026). Partisan prejudice: The role of beliefs about the unchanging nature of ideology and partisans. *Personality and Social Psychology Bulletin*, *52*(5), 1356–1385. https://doi.org/10.1177/01461672241283862

Huber, G. A., & Malhotra, N. (2017). Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics*, *79*(1), 269–283. https://doi.org/10.1086/687533

Hürter, T. (2024, December 31). Einführung der Gurtpflicht: Darf der Staat Autofahrer zum Überleben zwingen? *Die Zeit*. https://www.zeit.de/zeit-wissen/2025/01/einfuehrung-gurtpflicht-1976-freiheit-sicherheit-verkehr

Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological Science*, *7*(5), 496–503. https://doi.org/10.1177/1745691612448792

Insko, C. A., Wildschut, T., & Cohen, T. R. (2013). Interindividual–intergroup discontinuity in the prisoner's dilemma game: How common fate, proximity, and similarity affect intergroup

competition. *Organizational Behavior and Human Decision Processes*, *120*(2), 168–180. https://doi.org/10.1016/j.obhdp.2012.07.004

Jansson, F., & Strimling, P. (2025). Susceptibility to moral arguments among liberals and conservatives. *Public Opinion Quarterly*, *89*(3), 735–757. https://doi.org/10.1093/poq/nfaf045

Jia, C., Lam, M. S., Mai, M. C., Hancock, J. T., & Bernstein, M. S. (2024). Embedding democratic values into social media AIs via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–36. https://doi.org/10.1145/3641002

Kashima, Y., Kashima, E., Chiu, C.-Y., Farsides, T., Gelfand, M., Hong, Y.-Y., Kim, U., Strack, F., Werth, L., Yuki, M., & Yzerbyt, V. (2005). Culture, essentialism, and agency: Are individuals universally believed to be more real entities than groups? *European Journal of Social Psychology*, *35*(2), 147–169. https://doi.org/10.1002/ejsp.237

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, *55*(2), 187–195. https://doi.org/10.1037/0022-3514.55.2.187

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, *4*(3), 279–286. https://doi.org/10.1038/s41562-019-0766-4

Levendusky, M. S., & Malhotra, N. (2016). (Mis)perceptions of partisan polarization in the American public. *Public Opinion Quarterly*, *80*(S1), 378–391. https://doi.org/10.1093/poq/nfv045

Levendusky, M. S., & Stecula, D. A. (2023). *We need to talk: How cross-party dialogue reduces affective polarization*. Cambridge University Press. https://doi.org/10.1017/9781009042192

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, *78*(2), 223–246. https://doi.org/10.1037/0022-3514.78.2.223

Ma, N., Vandekerckhove, M., Van Overwalle, F., Seurinck, R., & Fias, W. (2011). Spontaneous and intentional trait inferences recruit a common mentalizing network to a different degree: Spontaneous inferences activate only its core areas. *Social Neuroscience*, *6*(2), 123–138. https://doi.org/10.1080/17470919.2010.485884

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327. https://doi.org/10.1177/1745691614568482

Mackie, D. M., & Queller, S. (2000). The impact of group membership on persuasion: Revisiting "who says what to whom with what effect?" In D. J. Terry & M. A. Hogg (Eds.), *Attitudes, behavior, and social context: The role of norms and group membership* (pp. 135–155). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410603210-8

Markant, D., Sah, S., Karduni, A., Rogha, M., Thai, M., & Dou, W. (2026). Correcting misperceptions at a glance: Using data visualizations to reduce political sectarianism. *IEEE Transactions on Visualization and Computer Graphics*, *32*(1), 1361–1371. https://doi.org/10.1109/TVCG.2025.3634777

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

McCoy, J., Rahman, T., & Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, *62*(1), 16–42. https://doi.org/10.1177/0002764218759576

Mitchell, J. P., Cloutier, J., Banaji, M. R., & Macrae, C. N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience*, *1*(1), 49–55. https://doi.org/10.1093/scan/nsl007

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*, *117*(26), 14864–14872. https://doi.org/10.1073/pnas.2001263117

Navon, M., Ratliff, K. A., & Corneille, O. (2026). Individuating rather than group information dominates evaluations of members from newly learned social groups. *Journal of Experimental Social Psychology*, *123*, Article 104851. https://doi.org/10.1016/j.jesp.2025.104851

Newman, L. S., & Uleman, J. S. (1990). Assimilation and contrast effects in spontaneous trait inference. *Personality and Social Psychology Bulletin*, *16*(2), 224–240. https://doi.org/10.1177/0146167290162004

Novoa, G., Echelbarger, M., Gelman, A., & Gelman, S. A. (2023). Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, *120*(47), Article e2309361120. https://doi.org/10.1073/pnas.2309361120

Padgett, J., Dunaway, J. L., & Darr, J. P. (2019). As seen on TV? How gatekeeping makes the U.S. house seem more extreme. *Journal of Communication*, *69*(6), 696–719. https://doi.org/10.1093/joc/jqz039

Paluck, E. L., Green, S. A., & Green, D. P. (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy*, *3*(2), 129–158. https://doi.org/10.1017/bpp.2018.25

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), Article 2333. https://doi.org/10.1038/s41467-022-30073-5

Pereira, A., & Van Prooijen, J.-W. (2018). Why we sometimes punish the innocent: The role of group entitativity in collective punishment. *PLOS ONE*, *13*(5), Article e0196852. https://doi.org/10.1371/journal.pone.0196852

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Quinn, K. A., Mason, M. F., & Macrae, C. N. (2009). Familiarity and person construal: Individuating knowledge moderates the automaticity of category activation. *European Journal of Social Psychology*, *39*(5), 852–861. https://doi.org/10.1002/ejsp.596

Rasinski, K. A., Crocker, J., & Hastie, R. (1985). Another look at sex stereotypes and social judgments: An analysis of the social perceiver's use of subjective probabilities. *Journal of Personality and Social Psychology*, *49*(2), 317–326. https://doi.org/10.1037/0022-3514.49.2.317

Reagans, R. (2011). Close encounters: Analyzing how social similarity and propinquity contribute to strong network connections. *Organization Science*, *22*(4), 835–849. https://doi.org/10.1287/orsc.1100.0587

Reinero, D. A., Harris, E. A., Rathje, S., Duke, A., & Van Bavel, J. J. (2023). Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims. *PsyArXiv*. https://doi.org/10.31234/osf.io/z4df3

Robbett, A., Colón, L., & Matthews, P. H. (2023). Partisan political beliefs and social learning. *Journal of Public Economics*, *220*, Article 104834. https://doi.org/10.1016/j.jpubeco.2023.104834

Roth, P. L., Arnold, J. D., Walker, H. J., Zhang, L., & Van Iddekinge, C. H. (2022). Organizational political affiliation and job seekers: If I don't identify with your party, am I still attracted? *Journal of Applied Psychology*, *107*(5), 724–745. https://doi.org/10.1037/apl0000932

Rubinstein, R. S., Jussim, L., & Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology*, *75*, 54–70. https://doi.org/10.1016/j.jesp.2017.11.009

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., Barea-Arroyo, P., Berge, M. L., Bjørndal, L. D., Bursalıoğlu, A., Bühler, V., Čadek, M., Çetinçelik, M., Clay, G., Cortijos-Bernabeu, A., Damnjanović, K., Dugue, T. M., Esberg, M., Esteban-Serna, C., … Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour*, *5*(10), 1369–1380. https://doi.org/10.1038/s41562-021-01092-x

Shafranek, R. M. (2021). Political considerations in nonpolitical decisions: A conjoint analysis of roommate choice. *Political Behavior*, *43*(1), 271–300. https://doi.org/10.1007/s11109-019-09554-9

Shehata, A. (2014). Game frames, issue frames, and mobilization: Disentangling the effects of frame exposure and motivated news attention on political cynicism and engagement. *International Journal of Public Opinion Research*, *26*(2), 157–177. https://doi.org/10.1093/ijpor/edt034

Simas, E. N., Clifford, S., & Kirkland, J. H. (2020). How empathic concern fuels political polarization. *American Political Science Review*, *114*(1), 258–269. https://doi.org/10.1017/S0003055419000534

Skowronski, J. J. (2002). Honesty and intelligence judgments of individuals and groups: The effects of entity-related behavior diagnosticity and implicit theories. *Social Cognition*, *20*(2), 136–169. https://doi.org/10.1521/soco.20.2.136.20993

Small, M. L., & Adler, L. (2019). The role of space in the formation of social ties. *Annual Review of Sociology*, *45*(1), 111–132. https://doi.org/10.1146/annurev-soc-073018-022707

Solanki, P., & Cesario, J. (2025). Stereotypes as Bayesian prediction of social groups. *The Journal of Social Psychology*, *165*(5), 640–662. https://doi.org/10.1080/00224545.2024.2368017

Spencer-Rodgers, J., Hamilton, D. L., & Sherman, S. J. (2007). The central role of entitativity in stereotypes of social categories and task groups. *Journal of Personality and Social Psychology*, *92*(3), 369–388. https://doi.org/10.1037/0022-3514.92.3.369

Thomas, R. J. (2019). Sources of friendship and structurally induced homophily across the life course. *Sociological Perspectives*, *62*(6), 822–843. https://doi.org/10.1177/0731121419828399

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, *83*(5), 1051–1065. https://doi.org/10.1037/0022-3514.83.5.1051

Törnberg, P., Andersson, C., Lindgren, K., & Banisch, S. (2021). Modeling the emergence of affective polarization in the social media society. *PLOS ONE*, *16*(10), Article e0258259. https://doi.org/10.1371/journal.pone.0258259

Trope, Y., & Liberman, A. (1996). Social hypothesis testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 239–270). The Guilford Press.

Uleman, J. S., Newman, L., & Winter, L. (1992). Can personality traits be inferred automatically? Spontaneous inferences require cognitive capacity at encoding. *Consciousness and Cognition*, *1*(1), 77–90. https://doi.org/10.1016/1053-8100(92)90049-G

Van Dessel, P., Ratliff, K., Brannon, S. M., Gawronski, B., & De Houwer, J. (2021). Illusory-correlation effects on implicit and explicit evaluation. *Personality and Social Psychology Bulletin*, *47*(10), 1480–1494. https://doi.org/10.1177/0146167220977706

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, *10*(2), 145–158. https://doi.org/10.1177/1745691615569849

Wojcieszak, M., & Warner, B. R. (2020). Can interparty contact reduce affective polarization? A systematic test of different forms of intergroup contact. *Political Communication*, *37*(6), 789–811. https://doi.org/10.1080/10584609.2020.1760406

Wolf, L. J., & Hanel, P. H. P. (2024). Correcting misperceptions of fundamental differences between U.S. Republicans and Democrats: Some hope-inspiring effects. *Social Psychological and Personality Science*, *15*(8), 895–907. https://doi.org/10.1177/19485506241263887

Wu, S., & Resnick, P. (2021). Cross-partisan discussions on YouTube: Conservatives talk to liberals but liberals don't talk to conservatives. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 808–819. https://doi.org/10.1609/icwsm.v15i1.18105

Yudkin, D. A., Hawkins, S., & Dixon, T. (2019). *The perception gap: How false impressions are pulling americans apart*. More in Common. https://doi.org/10.31234/osf.io/r3h5q

SUPPLEMENTARY MATERIALS I

First Article

Table of Contents

S1: Material Generation and Pretesting 145

S1.1: Stimulus Set 1 151

S1.2: Stimulus Set 2 155

S1.3: Filler Statements for Experiment 2 161

S1.4: Filler Statements for Experiment 4 163

S1.5: Filler Statements for Experiment 5 165

S2: Manipulation Checks 166

S2.1: Results of Manipulation Checks 167

S3: Additional Ratings of Materials 167

S3.1: Experiment 2 (Identity, Valence, & Accessibility) 167

S3.2: Experiment 4 (Identity, Necessity, & Base Rate) 167

S4: Collection of Demographic Data and Debriefing 168

S4.1: Political Orientation, Interest, and Satisfaction 169

S5: Additional Analyses 169

S5.1: By-Item Analyses Experiment 1 169

S5.2: By-Item Analyses Experiment 3 169

S5.3: Multiverse Analysis (Experiment 2) 170

S5.4: Error Rates (Experiment 2) 171

S5.5: Dislike and Distrust (Experiment 3) 171

S5.6: Multiverse Analysis (Experiment 4) 172

S5.7: Error Rates (Experiment 4) 173

S5.8: Order of Presentation (Experiment 4) 173

S5.9: Stimulus Set Version (Experiment 5) 173

References 174

S1: Material Generation and Pretesting

For the purposes of generating and validating our stimulus materials, we conducted a series of six pretests. Participants for all pretests were German-speaking users of the online platform Prolific (www.prolific.co). All pretests were run using Qualtrics (www.qualtrics.com).

In the first pretest we explained the concept of ideological categories to a sample of *N* = 20 participants (as "labels that people use to classify each other when dealing with each other's political or societal actions"). We listed the four examples *Feminist* (feminist), *Querdenker* (covidiot), *homophob* (homophobic), and *politisch korrekt* (politically correct) and asked participants to generate at least 10 further labels (noun or adjective) that they either use themselves or know other people to use. From this procedure resulted 213 categories. We excluded categories with defining criteria that we considered to be too clearly defined (e.g., *Black*), that are usually not used as person characteristics (e.g., *antifa*), that we deemed not ideological (e.g., *perfectionist*), too unusual (e.g., *colonialist*), or too general (e.g., *radical*) and added 13 categories that we had come up with ourselves—resulting in a set of 100 ideological categories that we submitted to the second pretest.

The aim of the second pretest was to generate behaviors or statements indicative of any of these 100 ideological categories. We presented a subset of 10 randomly drawn categories from the first pretest to each of *N* = 100 participants. They were instructed to generate three behaviors or statements per category from which they or other people might infer the presented category. They were further instructed focus on behaviors/statements that were a) as concrete as possible, b) frequently encountered in real life, and c) very specific to the respective category. The participants were allowed to skip as many categories as they wanted if they either did not know the category or strongly opposed its usage. On the basis of the participants' responses and additional material found by searching for the categories on social media platforms or in online news, we generated up to six behaviors/statements per category that met our requirements. This resulted in 274 behaviors/statements that were submitted to the third pretest.

SUPPLEMENTARY MATERIALS I

In the third pretest we tested which specific ideological categories were most consistently mentioned when asking the participants directly to infer ideological categories from the 274 generated behaviors/statements. We presented subsets of 50 randomly drawn behaviors/statements to each of 110 participants (resulting in 18 to 21 participants per behavioral description). The statements were presented in quotation marks whereas the behaviors were phrased in the third person with the actor referred to as "X" using both male and female pronouns (e.g., "In the canteen, everyone ordered a burger for lunch, but X ate his/her green spelt salad that he/she had brought from home."). The participants were asked to write down at least one and up to three ideological categories that spontaneously came to their mind when reading the behavior. To illuminate the concept of ideological category, we provided three exemplar labels, randomly drawing out of a set of ten. These were chosen so that two of the labels were nouns and one was an adjective and that there was little semantic overlap between them. The participants were instructed not to think about whether they thought it legitimate to label the person that way and to respond spontaneously. They were further instructed not to write down evaluations of the behavior/statement (such as good, reasonable, stupid, annoying, or dumb), but to focus on ideological categorizations.

Afterwards the participants were asked to state their own political attitudes using the one-item Left-Right Self-Placement scale (Breyer, 2015), their personal interest in politics using a one-item scale (Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V., 2014), and their satisfaction with Germany's political system using the German Satisfaction with the Political System Short Scale (SPS; Dentler, Bluemke, & Gabriel, 2020). Ultimately, they were debriefed and asked for their informed consent. The sample had a median age of 27 (ranging from 18 to 56). 58 participants identified as female, 47 as male and 5 as other. The participants reported to be rather left leaning (M = 3.81, SD = 1.75, on a scale ranging from 1 = "left" to 10 = "right"), rather interested in politics (M = 7.14, SD = 1.72, on a scale ranging from 1 = "not at

all" to 10 = "very strongly"), and rather satisfied with the political system (*M* = 2.39, *SD* = 0.51, on a scale ranging from 1 = "satisfied" to 4 = "dissatisfied").

We preprocessed the text data by converting it to lower case and removing all non-word characters. We then created a list with all unique category labels (*n* = 1797) and defined a list of replacements for labels that either were misspelled, had an uncommon declination, or were compound words containing a more common label. These simplified labels retained at least part of the morphology of the original label. We then clustered labels into morphologically distinct but semantically related labels using the most common labels as the cluster names (e.g., "sexist", "macho", and "patriarch" were clustered as "sexist"). The definition of the clusters was guided partly by researcher's intuition and partly by (visual) inspection of word correlations within the pretest data. For each behavior/statement we then calculated the consensus of the most common label by dividing the number of occurrences by the number of participants. We did this for both the simplified and the clustered labels independently. We then selected the 150 behaviors/statements with the highest consensus in the clustered labels (ranging from 25,0% to 95,2%) to be used for the next pretest. We used the more specific simplified labels as the implied ideological categories in order to retain potential subtle semantic differences.

From these 150 behaviors/statements we built stimulus set 1 (used in Experiments 1 and 2) by selecting 24 behaviors/statements that were most strongly associated with ideological categories (while making sure that they were semantically distinct). To complete the behavioral descriptions we added names, which were chosen such that the age and gender associated with the name did not seem incongruent with the respective behavior/statement. See **S1.1: Stimulus Set 1** for a table containing all behaviors, implied and implied-other ideological categories, consensus scores and *t* tests results.

For stimulus set 2 (used in Experiments 3, 4, and 5) we needed to create additional statements providing alternative reasons for each behavior. We therefore conducted a fourth pretest in which we presented randomly drawn subsets of 15 out of 150 behaviors/statements

together with the respective implied ideological category to each of *N* = 50 participants. We told them that participants of another study had generated these labels as explanations of the behaviors/statements and instructed them to come up with *more profound reasons* by thinking about experiences the actor might have had or about the situation the actor might be in. We explicitly told them that the reasons did not have to legitimize the behavior/statement or prove that the label does not apply to the actor. We furthermore asked them to come up with reasons that were not too unusual. Ultimately participants were asked how they had felt during the task, debriefed, and asked for their informed consent.

In response to the question about how they had felt during the task, 75% of the participants reported, that it was at least partly difficult to come up with reasons. 19% reported that it was particularly difficult for immoral behaviors and 15% reported that they felt by providing reasons they were legitimizing, justifying, or condoning the behaviors/statements. We used the provided reasons as a basis for generating the final materials, constructing three reasons per behavior/statement that did not explicitly contain mental states explaining the behavior (i.e., thoughts, feelings, goals) but from which such mental states could be inferred. We furthermore tried to construct reasons that by themselves did neither contradict nor support the attribution of the implied ideological label. We excluded one statement because we thought the assigned category did not really fit the statement and three behaviors/statements because they explicitly contained the label (we had not noticed this before). This yielded a total of 146 behaviors/statements and 3 * 146 = 438 reasons.

In the fifth pretest we wanted to select the best out of three reasons constructed for each of the 146 behaviors/statements. In order to do this, we presented subsets of 36 randomly drawn behaviors/statements together with one of their respective reasons to each of *N* = 183 participants. We asked them to rate how well each reason (which we referred to as “background information”) could explain the corresponding behavior/statement on a 5-point scale (ranging from “very badly” to “very well”). Because all the reasons were intended to be good explanations, we also included four filler items with bad reasons drawn randomly from a pool of ten items. The filler items were

presented at fixed positions (2nd, 7th, 17th, and 28th). We constructed three parallel versions of the task so that no participant saw the same behavior/statement more than once. Participants were randomly assigned to one of these three versions.

In the second task of the pretest, the aim was to validate the implied ideological categories. Thus, we presented subsets of 12 randomly drawn behaviors/statements together with their respective ideological category to the participants. We again asked them to rate how well each category (referred to as "trait") could explain the corresponding behavior/statement, using the same scale as in the first task. We inserted three filler items with incongruent categories drawn randomly from a pool of ten items. The filler items were presented at fixed positions (2nd, 3rd, and 10th). The behaviors/statements presented in this task had not yet been encountered in the first task. After the second task we asked the participants to indicate which task they considered easier on a 5-point scale (ranging from "the first one was considerably easier" to "the second one was considerably easier"). We then asked them to write as much as they like on whether they felt differently during the two tasks. Ultimately participants were debriefed and asked for their informed consent.

Participants rated the second task as easier than the first task (M = 0.99, SD = 1.10, on a scale ranging from −2 = to 2). For each behavior we selected the reason with the highest average score. In a few cases we used the second highest scoring reason because the difference was negligible and we had a subjective preference for the lower scoring reason. We thus selected 48 behaviors/statements that had high ratings both regarding the reason as well as the ideological category. For these 48 behaviors/statements we constructed control versions of the reasons containing more or less the same words but without implying the same explanation of the behavior/statement.

The aim of the sixth and final pretest was to identify behaviors/statements with the largest difference between reason and control version. We thus presented the 48 behaviors/statements together with the reasons or control versions to 34 participants. We asked them to rate whether each ancillary statement was a sufficient explanation of the behavior on a 5-point scale (ranging

from “no” to “yes”). We constructed two parallel versions of the task so that no participant saw the same behavior/statement more than once. Participants were randomly assigned to one of these two versions. Because we expected some participants to respond randomly, we excluded participants from the analysis whose ratings of the reasons and control reasons did not significantly differ from one another at a level of α = .05. This led to the exclusion of 4 participants. We sorted the behaviors/statements by the size of the difference between the ratings of the reason and control version. We constructed a set of 24 behaviors/statements (stimulus set 2) by successively including behaviors from the top and skipping behaviors which implied categories that were synonymous to those already included. See **S1.2: Stimulus Set 2** for a table containing all behaviors, reasons, control reasons, implied and implied-other ideological categories, consensus scores, reason scores, category scores, and *t* tests results. Materials in the original language (German) can be found in the osf-project (https://osf.io/f5bv9/).

S1.1: Stimulus Set 1

Table S1.1

Stimulus Set 1 with Pretest-Results and Inference-Effects of Experiment 1

		Pretest Results		Inference-Effects (Experiment 1)					
Category Implied (-Other)	**Behavior**	**Category Score**	**Consensus Simplified (Clustered)**	**Implied-other *M*(*SD*)**	**Implied *M*(*SD*)**	***t***	***df***	***p***	***d*$_z$**
AfD member (leftist)	Michael says: "All the asylum seekers only come to Germany to get our social benefits. We shouldn't make it so easy for them, too."	0.94	32% (32%)	1.26 (0.66)	4.07 (0.78)	14,61	26	<.001	2,81
atheist (esoteric)	Sarah has an appointment today at the Citizens Registration Office to withdraw from the church.	1.44	38% (43%)	1.48 (0.58)	3.67 (1.21)	8,53	26	<.001	1,64
esoteric (covidiot)	Sonja pours water for her guests from a pitcher that has little colorful stones at the bottom.	1.21	35% (35%)	2.11 (0.93)	3.52 (1.09)	7,53	26	<.001	1,45
feminist (atheist)	Monika says, "Women still don't really have equal rights. The ongoing gender pay gap alone shows that."	1.29	70% (85%)	2.11 (0.93)	4.22 (0.75)	10,1	26	<.001	1,94
green (sexist)	Christina's bicycle bears a "Nuclear power? No thanks" sticker.	1.50	47% (74%)	1.11 (0.32)	4.30 (0.82)	18,84	26	<.001	3,62
hippie (conservative)	Vanessa says, "You have to mentally break away from all this striving for achievement and focus on what's really important: peace, love, nature."	1.69	68% (68%)	1.96 (0.90)	3.63 (0.97)	8,06	26	<.001	1,55

homophobic (liberal)	Nina says, “I don’t mind them doing what they want in their own homes, but when I’m at the park with my kids, I don’t want to see two men sticking their tongues down each other’s throats.”	1.71	65% (65%)	1.74 (0.71)	4.15 (1.06)	7,71	26	<.001	1,48
incel (punk)	Marvin says, “If even completely unattractive women only chase ripped models, it’s obvious why I’ll never have a girlfriend.”	0.71	30% (30%)	1.33 (0.48)	3.85 (1.06)	11,66	26	<.001	2,24
climate change denier (homophobic)	Ralf says: “There have always been warmer and colder phases in the history of the earth. But that doesn’t mean we have to imagine the worst.”	1.94	24% (62%)	1.56 (0.97)	4.44 (0.70)	16,08	26	<.001	3,09
communist (conspiracy theorist)	Maike says, “Financial capitalism has had its day as a system and must finally be overturned.”	1.50	40% (50%)	2.00 (0.73)	3.44 (0.89)	7,15	26	<.001	1,38
conservative (climate change denier)	Marion says, “In the good old days, people could still speak normally. What people speak today doesn’t sound like German at all anymore with all the gender gaps and anglicisms.”	1.18	58% (58%)	2.26 (1.06)	4.41 (0.89)	7,51	26	<.001	1,45
liberal (racist)	Simone says, “Our society is far too bureaucratic. I’m in favor of a leaner state and more personal responsibility.”	1.31	35% (35%)	1.78 (0.85)	4.15 (0.82)	11,41	26	<.001	2,2
leftist (nazi)	Katharina was at a demonstration today against foreign deployments of the German Armed Forces.	1.22	45% (45%)	1.74 (1.02)	3.41 (0.97)	4,94	26	<.001	0,95
nazi (green)	Heike says: “It should be allowed to say that not everything was bad in the Third Reich.”	1.38	40% (40%)	1.52 (0.70)	3.44 (1.19)	6,11	26	<.001	1,17

pacifist (AfD member)	Nico says: “The EU should be demilitarized. If we don’t have any weapons, then other nations will also disarm and there will be an end to pointless wars.”	1.67	29% (38%)	1.41 (0.57)	4.48 (0.80)	17,42	26	<.001	3,35
punk (religious)	Emil has colored hair and is dressed completely in black. His sweater has ACAB written on it in big letters.	0.69	50% (50%)	1.70 (0.67)	4.18 (1.04)	10,56	26	<.001	2,03
covidiot (communist)	Susanna says: “We live in a health dictatorship! Our basic rights have been taken away from us for over a year!”	1.24	65% (65%)	1.74 (0.71)	3.89 (1.05)	8,64	26	<.001	1,66
racist (hippie)	Barbara says, “I hope the new colleague will finally be a German again.”	1.29	67% (67%)	1.44 (0.64)	4.00 (1.00)	11,85	26	<.001	2,28
religious (trumpist)	Claudia was insanely exhausted from the long day but still took ten minutes to pray before going to bed.	1.93	61% (89%)	1.81 (0.83)	4.89 (0.42)	18,28	26	<.001	3,52
sexist (pacifist)	When Jannik’s girlfriend doesn’t dare to jump from the three-meter board, he tells her not to be a pussy.	0.77	40% (75%)	1.96 (0.94)	3.33 (1.49)	4,05	26	<.001	0,78
transphobic (vegan)	Natalie says, “Trans women shouldn’t be allowed to use women’s restrooms. Women need places where they are safe from male assault and not every disguised man can walk in.”	1.27	40% (40%)	1.41 (0.64)	4.22 (1.12)	12,09	26	<.001	2,33
trumpist (incel)	John is wearing a cap with the words “Make America Great Again” on it.	1.75	48% (48%)	2.30 (1.14)	4.63 (0.69)	10,03	26	<.001	1,93
vegan (transphobic)	Larissa asks the waitress at the restaurant if the mixed salad contains any animal products.	1.58	90% (95%)	1.07 (0.27)	4.44 (0.58)	27,83	26	<.001	5,36

conspiracy theorist (feminist)	Sina says, “I do believe that we were not told the whole story about 9/11. The terrorist attack was definitely in the interest of the American government.”	1.38	65% (70%)	1.96 (0.85)	3.93 (1.03)	8,55	26	<.001	1,65

Note. Category labels and behaviors were translated from German to English. The original versions can be found in the osf-project. The category score was derived from the fifth pretest and ranges from −2 to 2 with higher values indicating that the ideological categories explained the behaviors/statements well. Consensus data is derived from the third pretest by dividing the number of occurrences of the most common ideological category by the number of participants. For the simplified consensus we pooled category labels such that they retained at least part of their morphology. For the clustered consensus we also pooled synonyms that were morphologically distinct. For the inference effects (Experiment 1) we calculated means and standard deviations (the scale ranged from 1 to 5) and ran one-tailed dependent samples *t* tests for each item separately. We report Cohen’s d_z for all items.

S1.2: Stimulus Set 2

Table S1.2

Stimulus Set 2 with Pretest-Results and Discounting-Effects of Experiment 3

Discounting-Effects (Experiment 3)	**Pretest Results**	**Category Implied (-Other)**	religious (homophobic)	vegan (racist)
	Ancillary Statement	**Behavior**	Claudia was insanely exhausted from the long day but still took ten minutes to pray before going to bed.	Michelle decides not to buy coffee after all, after the waiter said there was no soy milk.
		Reason	In the past, Claudia often had problems falling asleep. Since she started praying in the evening, she can fall asleep much better.	Michelle saw a gastroenterologist a few weeks ago and learned that her digestive problems are due to lactose intolerance.
		Control	In the past, Claudia often had problems falling asleep. Since she started avoiding caffeine in the evening, she can fall asleep much better.	Michelle' s friend saw a gastroenterologist the day before yesterday and was told that his digestive problems are not due to lactose intolerance.
		Reason Difference Score	2.93	2.87
		Category Score	1.93	1.56
		Consensus Simplified (Clustered)	61% (89%)	60% (65%)
		Control *M* (*SD*)	2.61 (1.08)	1.85 (1.42)
		Reason *M* (*SD*)	2.55 (1.24)	0.89 (1.18)
		t	0.36	5.21
		df	194	192
		p	.716	<.001
		d	0.05	0.74

homophobic (pacifist)	transphobic (vegan)	atheist (transphobic)	feminist (capitalist)
Nina says, " I don' t mind them doing what they want in their own homes, but when I' m at the park with my kids, I don' t want to see two men sticking their tongues down each other' s throats."	Isabell says, " If trans women are now allowed to compete against us real women in sports, then we can get right back behind the stove."	Sarah has an appointment today at the Citizens Registration Office to withdraw from the church.	Monika says, " Women still don' t really have equal rights. The ongoing gender pay gap alone shows that."
Nina was just taking a walk in the park with her family and saw two men having sex in the bushes.	Isabell just lost for the third time in a row to a trans woman from her tennis club.	Recently, it was revealed by some of the victims that there had also been cases of abuse in Sarah' s church congregation.	Monika' s daughter and her husband have the same job in the same company, but he earns almost twice as much as she does.
While taking a walk with her family, Nina remembered that in this park she already had sex in the bushes with two men.	The trans woman from Isabell' s tennis club has just lost to Isabell for the third time in a row.	Recently, Sarah' s friend told her that her church congregation was raising money for victims of abuse.	Monika' s daughter and her husband have the same job in the same company, and each earn almost twice as much as Monika.
2.73	2.60	2.53	2.40
1.71	0.53	1.44	1.29
65% (65%)	38% (38%)	38% (43%)	70% (85%)
2.98 (1.16)	2.46 (1.42)	1.62 (1.56)	1.45 (1.45)
1.52 (1.30)	1.50 (1.42)	1.36 (1.32)	1.80 (1.24)
8.37	4.78	1.27	−1.83
196	198	193	193
<.001	<.001	.205	.068
1.18	0.68	0.18	−0.26

covidiot (climate change denier)	snob (green)	nazi (snob)	conservative (communist)
Juliane says, “ I am worried about the vaccination, especially with my children. Who knows the damage these genetically engineered vaccines might cause.”	Moritz says: “ Football is a bit too proletarian for me; I really can’ t watch it. I’ d rather watch a tennis match.”	Jutta says: “ Hitler did a lot of good for Germany. Some of it would be necessary again these days.”	Bernd says, “ People who want to live in Germany also have to adapt to our cultural values.”
The alternative practitioner who helped Juliane with her irritable bowel syndrome said that mRNA vaccines are dangerous and alter the genome.	Moritz was once taken to a football game by friends and then almost beaten up by fans of the opposing team.	Her grandfather told Jutta that Hitler reduced the unemployment rate significantly, which made their family relatively prosperous.	Bernd’ s sister is a teacher and says that the students with a Turkish migratory background often don’ t take her or other teachers seriously.
The alternative practitioner who helped Juliane with her irritable bowel syndrome asserts that mRNA vaccines are harmless and do not alter the genome.	Moritz was taken to a football game by friends. The opposing team’ s fans clapped their hands.	Jutta’ s grandfather told her that Hitler greatly reduced his family’ s wealth and caused a high unemployment rate.	Bernd’ s sister is a teacher and says that the other teachers often don’ t take the students with a Turkish migratory background seriously.
2.13	2.07	2.00	1.93
1.07	0.93	1.50	0.88
45% (45%)	20% (35%)	68% (68%)	45% (45%)
0.90 (1.09)	1.08 (1.60)	1.38 (1.56)	2.33 (1.26)
1.03 (1.21)	0.60 (1.66)	0.51 (1.50)	2.21 (1.30)
−0.80	2.08	4.02	0.66
196	198	198	198
.425	.039	<.001	.508
−0.11	0.29	0.57	0.09

racist (atheist)	hippie (queer)	communist (conservative)	leftist (incel)
Astrid says, " You speak good German for the way you look. Where are you from?"	Vanessa says, " You have to mentally break away from all this striving for achievement and focus on what' s really important: peace, love, nature."	Maike says, " Financial capitalism has had its day as a system and must finally be overturned."	Björn says, " The wealth gap is opening up far too much. If ordinary workers can' t feed their families anymore, then capitalism has obviously failed."
Astrid has lived her whole life in a tiny village and has never met any non-white people whose mother tongue was German.	Vanessa' s grandfather was one of the most important people in her life. On his deathbed, he deeply regretted having dedicated his entire life to his career.	Due to the 2008 financial crisis, Maike lost a lot of money that she had previously invested in stocks.	Björn and his wife work full-time in a workshop. Although they are expecting a child, they have to move to a smaller apartment for financial reasons.
Astrid doesn' t come from a tiny village but from a big city, which is also home to many non-white people whose mother tongue is German.	Vanessa' s grandfather was one of the most important people in her life. On his deathbed, he deeply regretted not having fully supported her in her career.	Although Maike had invested a lot of money in stocks, she hardly lost anything due to the 2008 financial crisis.	Björn and his wife are expecting a child. Since Björn is now working full-time in the workshop, they could afford an apartment with a small garden.
1.93	1.87	1.80	1.80
1.12	1.69	1.50	1.08
70% (70%)	68% (68%)	40% (50%)	55% (55%)
1.31 (1.56)	0.81 (1.19)	0.070 (1.65)	1.76 (1.33)
0.94 (1.35)	0.67 (1.16)	0.15 (1.71)	1.85 (1.23)
1.79	0.85	−0.34	−0.50
194	198	198	197
.075	.399	.737	.620
0.25	0.12	−0.05	−0.07

conspiracy theorist (radical leftist)	capitalist (islamophobic)	radical leftist (covidiot)	Islamist (pacifist)
Sina says, " I do believe that we were not told the whole story about 9/11. The terrorist attack was definitely in the interest of the American government."	Mark says, " Government administration is simply far too inefficient to do its job properly. We should hand over as many tasks as possible to the private sector."	Timo says: " Are you coming to the May Day protest? I really feel like punching a few Nazis again."	Tarek says, " Islam is the only correct faith and must apply to all people."
A friend of Sina' s who is interested in politics explained to her that the US did not invade Iraq to fight terrorism but because of the oil.	Mark has read that the local health authorities are still sending the number of infections to the RKI via fax machine rather than digitally.	When he was young, an asylum seeker was killed by a group of right-wing extremists in Timo' s hometown.	A friend of Tarek' s took him to a sermon. The preacher said God would reward one for proselytizing unbelievers.
Sina talked to a friend from the US who is interested in politics about the fight against terrorism in Iraq. Because he couldn' t think of anything, he explained something about oil to her.	Mark' s mother sent him a report by the RKI and the local health authorities on the number of infections via fax instead of digitally.	When he was young, a right-wing extremist was killed by a group of asylum seekers in Timo' s hometown.	A friend of Tarek' s was taken to a sermon where the preacher said God does not reward one for proselytizing unbelievers.
1.53	1.47	1.40	1.33
1.38	1.33	1.38	0.92
65% (70%)	25% (35%)	29% (57%)	30% (30%)
1.45 (1.34)	1.93 (1.32)	2.07 (1.75)	1.33 (1.68)
1.22 (1.32)	1.94 (1.20)	2.28 (1.56)	1.32 (1.44)
1.22	−0.056	−0.89	0.045
198	196	195	193
.223	.955	.373	.964
0.17	−0.01	−0.13	0.01

pacifist (religious)	green (nazi)	islamophobic (feminist)	incel (hippie)
Nico says: " The EU should be demilitarized. If we don' t have any weapons, then other nations will also disarm and there will be an end to pointless wars."	Christina' s bicycle bears a " Nuclear power? No thanks" sticker.	Sascha says, " I don' t think it' s just the misinterpretation of the Koran. The religion is just inherently violent and that' s virtually impossible to change."	Marvin says, " If even completely unattractive women only chase ripped models, it' s obvious why I' ll never have a girlfriend."
When Nico was still a child, his father died during his military service.	Christina' s mother fell ill with thyroid cancer a few years after the reactor disaster in Chernobyl.	Sascha listened to an interview with Hamed Abdel-Samad, according to who the IS is doing exactly what Prophet Mohammed did and what is enshrined in the Koran.	Marvin downloaded Tinder a few days ago, swiped for hours and still hasn' t had a single match.
Nico was still a child when his father died just a few days before the start of his military service.	Christina' s mother had developed thyroid cancer a few years before the reactor disaster in Chernobyl.	Sascha heard an interview about Hamed Abdel-Samad, the Prophet Mohammed, the Koran and in the end also about the IS.	Marvin' s brother downloaded Tinder a few days ago, swiped for hours and got several matches.
1.27	1.20	1.13	1.07
1.67	1.50	1.07	0.71
29% (38%)	47% (74%)	16% (32%)	30% (30%)
1.98 (1.39)	2.83 (1.09)	1.28 (1.66)	1.69 (1.30)
1.96 (1.35)	3.01 (0.97)	1.92 (1.53)	1.81 (1.46)
0.10	−1.23	−2.83	−0.61
198	195	197	195
.918	.219	.005	.540
0.01	−0.17	−0.40	−0.09

queer (leftist)	climate change denier (conspiracy theorist)
Antonia's Instagram bio says, "trans women are women" and "she/her".	Ralf says: "There have always been warmer and colder phases in the history of the earth. But that doesn't mean we have to imagine the worst."
A friend of Antonia recently came out as trans to her.	Ralf talks to a young climate activist who tearfully says that if we don't reach the 1.5-degree target, the whole world will descend into chaos.
A friend of Antonia's is trans but hasn't come out to her yet.	Someone told Ralf about a young climate activist who tearfully said the world would descend into chaos if we didn't reach the 1.5-degree target.
0.80	0.80
1.13	1.94
21% (47%)	24% (62%)
−0.43 (1.59)	0.97 (0.94)
−0.66 (1.42)	0.92 (0.95)
1.08	0.37
196	198
.283	.708
0.15	0.05

Note. Category labels, behaviors, and ancillary statements were translated from German to English. The original versions can be found in the osf-project. The reason difference score was derived from the sixth pretest and calculated by subtracting the rating for the control statement from the rating of the reason. Difference scores range from −4 to 4 with higher values indicating that reasons more sufficiently explain the behaviors compared to control statements. The category score was derived from the fifth pretest and ranges from −2 to 2 with higher values indicating that the ideological categories explained the behaviors/statements well. Consensus data is derived from the third pretest by dividing the number of occurrences of the most common ideological category by the number of participants. For the simplified consensus we pooled category labels such that they retained at least part of their morphology. For the clustered consensus we also pooled synonyms that were morphologically. For the inference effects (Experiment 3) we subtracted the rating of the implied-other category from the rating of the implied category and ran two-sample *t* tests for each item separately. The resulting difference scores can range from −4 to 4 with higher values indicating stronger inference effects. We report Cohen's d for all items.

S1.3: Filler Statements for Experiment 2

Table S1.3

Filler Statements for Experiment 2

Behavior
Tarek knows many of his friends from an **incel** forum. However, some of them seem a bit extremist to him.
Sascha says, "I think as a **pacifist** you can only approve of the withdrawal from Afghanistan."
Patrick says, "I'm so glad my parents are **atheists** and not fundamental Christians."
Joachim says, "I don't think it's far-right that **AfD voters** want Germany and Germans to do well."
Moritz says, "I sort of believe homeopathy works, but I was also raised by a crazy **hippie** and an **esoteric**."
Mark went to a demonstration yesterday with a friend who is a **climate change denier**.
Isabell says, "I don't think it's **transphobic** to ask for trans women not to be allowed to participate in women's sports."
Jürgen says, "I'm tired of the politically correct lying press always defaming us as **covidiots** or **conspiracy theorists**."
Juliane says, "I think it's unpleasant to be berated right away as a **covidiot**. Who knows exactly what all these genetically modified vaccines do."
Bernd says: "Our children are really tiresome at the moment. You're constantly being accused of being **conservative**. Or **homophobic** or **transphobic** or whatever."
Sami's girlfriend thinks he's unfortunately a bit of a **conservative sexist**. He has never done the laundry since they started living together.
Antonia's Instagram bio reads "**liberal feminist**, **pacifist**, and devoted servant of two cats."
Leon's friends call him their little **green hippie** or **esoteric** because he always walks around the university barefoot.
Katrin actually found Donald Trump's policies quite convincing but never dared to come out as a **Trumpist** to her **liberal** friends.
Lars says, "I'd rather be a naive **left-wing communist** than join the ranks of these **racist** capitalists."
When Markus made a joke about a teammate after practice, the others said the joke was **sexist** and **homophobic**.
Björn says: "I don't understand why so few people vote for the **left**. Most would benefit from it, after all. You don't have to become a **communist** or a **punk**."

Timo says to his **punk** friends: “Are you coming to the First of May protest? I really feel like punching a few **Nazis** once again.”

Jasmin’s best friend is a dedicated activist and **feminist** and regularly takes her to workshops for FLINTA* people.

Since Michelle became a **vegan**, she drinks much less coffee because there is still no soy milk in the cafeteria.

Steffi says, “You really shouldn’t read Youtube comments, it’s full of **incels**, **Trumpists** and **conspiracy theorists**.”

Manuel says: “I think I have to move. One of my roommates is a **green vegan** and the other is a **religious climate denier**. It’s unbearable.”

Astrid says, “Now I’m an **AfD voter**, a **racist** or even a **Nazi** just because I asked someone where they come from?”

Matthias says: “I was never an **atheist**, but I have only recently become truly **religious**. Before, I just couldn’t understand God’s signs.”

Note. Behaviors were translated from German to English. The original versions can be found in the osf-project. Ideological categories serving as implied and implied-other probes in the test stimuli are printed in bold.

S1.4: Filler Statements for Experiment 4

Table S1.4

Filler Statements for Experiment 4

Behavior	Ancillary Statement
Jonas knows many of his friends from an **incel** forum. However, some of them seem a bit extremist to him.	A friend of Jonas says he would like to get revenge on all the women who have ever turned him down.
Emil says, “I think as a **pacifist** you can only approve of the withdrawal from Afghanistan.”	One of Emil’s classmates just told him about an interesting documentary about the U.S. intervention in Afghanistan.
Patrick says, “I’m so glad my parents are **atheists** rather than fundamental Christians.”	Patrick watched an interview with a woman whose strict Christian parents had forbidden her to go to college.
Joachim says, “I think **radical leftists** often behave very immorally, even though they may have good intentions.”	A colleague of Joachim’s said he had kind of rejoiced at the riots during the G20 protests.
Michael says, “I sort of believe homeopathy works, but I was also raised by a crazy **hippie**.”	Michael talks with a friend whose parents were quite the **snobs** about their family upbringing.

Thomas went to a protest yesterday with a friend who is a **climate change denier**.	Thomas' date, with whom he was supposed to meet yesterday, canceled on him at short notice.
Stefanie is a member of an association that is currently planning an action against **transphobic** discrimination.	Stefanie watches many documentaries and videos about trans people's experiences of discrimination.
Jürgen says, "I'm tired of the politically correct lying press always defaming us as **covidiots** or **conspiracy theorists**."	Jürgen read an article railing against those who attended a protest against the covid-19 measures.
Sabine says: "I think it's not nice to be branded as a **covidiot** right away. I simply don't want to be told what to do by the government."	Sabine's husband feels that the state abused its power during the pandemic.
Dirk says: "Our children are really exhausting right now. They're constantly accusing us of being **conservative**. Or **homophobic** or **transphobic** or whatever."	Dirk's daughter just stormed out of the room in a huff because she didn't like what he said.
Sami's girlfriend thinks he's unfortunately a bit of a **conservative capitalist**. He rants about the welfare state all the time.	Sami's parents always taught him that everyone is the architect of his own fortune.
In Anja's political science seminar, most people have very **queer**, **feminist**, and **pacifist** views.	Anja couldn't decide on a political science seminar and just picked one at random.
Leon's friends refer to him as their little **green hippie** because he always walks around the university barefoot.	Leon's **snobby** father has threatened to cut him off if he doesn't dress decently.
Katrin read in a blog that the oppression of **queer** people is inextricably linked to the class rule of **capitalists**.	Katrin had absolutely nothing to do at work today, so she spent hours surfing the internet.
Lars says, "I'd rather be a naive **leftist communist** than join the ranks of these **racist** capitalists."	Lars was just having a discussion with some people in a bar. They were very arrogant and treated him like a child.
When Markus made a joke about a teammate after practice, the others said his joke was **homophobic** and **islamophobic**.	The people on Markus' soccer team often make fun of the fact that he just can't tell jokes.
David says: "I don't understand why so few people vote for the **left**. After all, most would benefit from it. You don't have to become a **far-left communist**."	David is looking at the results of the state election with his girlfriend.
Nils says, "His statement was definitely **islamophobic**, but I wouldn't expect anything else from a **nazi** like him."	Nils' roommate just read him a tweet from an AfD politician.
Jasmin's best friend is a dedicated activist and **feminist** and regularly takes her to workshops for FLINTA* people.	Before she completed an insanely stressful internship, Jasmin always wanted to be a manager at a large company.

Sandra works at a hip start-up that produces protein shakes for **vegans**.	Sandra once had an allergic reaction to a protein shake and hasn't touched them since.
Jessica says, "You really shouldn't read YouTube comments, it's full of **incels**, **islamists**, and **conspiracy theorists**."	Since Jessica's sharing her YouTube account with her roommate, she's been suggested very bizarre videos.
Manuel says: "I think I have to move. One of my roommates is a **green vegan** and the other is a **climate change denier**. It's unbearable."	Manuel's sister says she'll turn religious if his nutty roommates ever get along.
Dagmar says, "So now I'm already a **racist** or even a **nazi**, just for saying that there will definitely be a few **Islamists** among all those Muslims?"	Dagmar spent her last summer vacation in Turkey.
Matthias says: "I was never an **atheist**, but I have only recently become truly **religious**. Before, I just couldn't understand God's signs."	Matthias' ex-girlfriend always tried to persuade him to come to church with her, but he didn't want to.

Note. Behaviors and ancillary statements were translated from German to English. The original versions can be found in the osf-project. Ideological categories serving as implied and implied-other probes in the test stimuli are printed in bold.

S1.5: Filler Statements for Experiment 5

Table S1.5

Filler Statements for Experiment 5

Behavior	Ancillary Statement
Jonas knows many of his friends from an **incel** forum. However, some of them seem a bit extremist to him.	A friend of Jonas says he would like to get revenge on all the women who have ever turned him down.
Thomas went to a protest yesterday with a friend who is a **climate change denier**.	Thomas' date, with whom he was supposed to meet yesterday, canceled on him at short notice.
Stefanie is a member of an association that is currently planning an action against **transphobic** discrimination.	Stefanie watches many documentaries and videos about trans people's experiences of discrimination.
Jürgen says, "I'm tired of the politically correct lying press always defaming us as **covidiots** or **conspiracy theorists**."	Jürgen read an article railing against those who attended a protest against the covid-19 measures.
Sami's girlfriend thinks he's unfortunately a bit of a **conservative capitalist**. He rants about the welfare state all the time.	Sami's parents always taught him that everyone is the architect of his own fortune.

In Anja's political science seminar, most people have very **queer**, **feminist**, and **pacifist** views.	Anja couldn't decide on a political science seminar and just picked one at random.
Leon's friends refer to him as their little **green hippie** because he always walks around the university barefoot.	Leon's **snobby** father has threatened to cut him off if he doesn't dress decently.
When Markus made a joke about a teammate after practice, the others said his joke was **homophobic** and **islamophobic**.	The people on Markus' soccer team often make fun of the fact that he just can't tell jokes.
David says: "I don't understand why so few people vote for the **left**. After all, most would benefit from it. You don't have to become a **far-left communist**."	David is looking at the results of the state election with his girlfriend.
Sandra works at a hip start-up that produces protein shakes for **vegans**.	Sandra once had an allergic reaction to a protein shake and hasn't touched them since.
Dagmar says, "So now I'm already a **racist** or even a **nazi**, just for saying that there will definitely be a few **Islamists** among all those Muslims?"	Dagmar spent her last summer vacation in Turkey.
Matthias says: "I was never an **atheist**, but I have only recently become truly **religious**. Before, I just couldn't understand God's signs."	Matthias' ex-girlfriend always tried to persuade him to come to church with her, but he didn't want to.

Note. Behaviors and ancillary statements were translated from German to English. The original versions can be found in the osf-project. Ideological categories serving as implied and implied-other probes in the test stimuli are printed in bold.

S2: Manipulation Checks

Experiments 3, 4, and 5 included manipulation checks. The manipulation check was completed after the main task (and in Experiment 4 after the collection of additional ratings). Participants rated the degree to which each presented ancillary statement provided a sufficient explanation of the respective behavior on a five-point Likert scale ranging from "no" to "yes". For the analysis we conducted one-tailed dependent samples *t* tests. In all three experiments we found large significant effects, indicating that participants rated the ancillary statements containing reasons to explain the behaviors more sufficiently than the ones without reasons (see Table S2.1).

S2.1: Results of Manipulation Checks

Table S2.1

Results for Manipulation Checks in Experiments 3, 4, and 5

	Reason		Control					
	M	***SD***	***M***	***SD***	***df***	***t***	***p***	d_z
Exp. 3	3.55	0.63	2.27	0.61	199	23.3	< .001	1.65
Exp. 4	3.48	0.57	2.25	0.58	235	26.3	< .001	1.72
Exp. 5	3.30	0.57	2.28	0.56	385	27.3	< .001	1.39

Note. Ratings of the ancillary statements were assessed on a Likert scale ranging from 1 = "no" to 5 = "yes" with higher rated statements being perceived as more sufficient explanations of the behavior.

S3: Additional Ratings of Materials

S3.1: Experiment 2 (Identity, Valence, & Accessibility)

In Experiment 2, after the probe recognition task, we collected additional ratings of the implied categories. Participants rated the valence of each category on a five-point scale ranging from "very negative" to "very positive", how much they identify with the category on a five-point scale ranging from "not at all" to "very strongly", and how often they use the category label on a five-point scale ranging from "never" to "very often". The order of the three ratings was counterbalanced across participants. The rating data can be found in the osf-project.

S3.2: Experiment 4 (Identity, Necessity, & Base Rate)

In Experiment 4, after the probe recognition task, we collected additional ratings of the implied categories. Participants rated how much they identify with the category on a five-point scale ranging from "not at all" to "very strongly", the degree to which the category is necessary to explain the respective behavior on a five-point scale ranging from "no" to "yes", and the base rate of the label on a five-point scale ranging from "very low" to "very high". The order of these three ratings is counterbalanced across participants. The rating data can be found in the osf-project.

S4: Collection of Demographic Data and Debriefing

Collection of demographic data and debriefing was similar in all five experiments. After the main task (and after giving additional ratings of the materials in Exp. 2 and 4) participants' own political orientation was assessed via the one-item Left-Right Self-Placement scale (Breyer, 2015). Next, their personal interest in politics was assessed via a one-item scale (Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V., 2014). In all but the last experiment they were also asked to state their satisfaction with Germany's political system via the German Satisfaction with the Political System Short Scale (SPS; Dentler, Bluemke, & Gabriel, 2020). Summary statistics of these measures can be found in Table S4.1. Ultimately, they indicated gender and age.

In all but the last experiment participants were also asked to rate on a 10-point scale how diligently they followed the instructions (ranging from "not at all" to "very conscientiously"). Furthermore, they were asked to indicate whether they thought their data was fit for analyses ("yes" or "no"). These questions served our preregistered exclusion criteria. Ultimately participants were asked whether they had any assumptions about the exact purpose of the experiment, debriefed about the purpose of the experiment, and asked for their informed consent for data storage and analysis.

S4.1: Political Orientation, Interest, and Satisfaction

Table S4.1

Political Orientation, Interest, and Satisfaction for All Samples

Variable	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Political orientation[a]					
M (*SD*)	3.5 (1.5)	3.9 (1.7)	4.1 (1.5)	3.8 (1.4)	3.9 (1.5)
Political interest[b]					
M (*SD*)	6.6 (2.5)	6.6 (2.2)	6.3 (2.2)	6.2 (2.1)	6.3 (2.0)
Political satisfaction[c]					
M (*SD*)	2.4 (0.5)	2.3 (0.5)	2.4 (0.5)	2.5 (0.5)	–

Note. [a] Political orientation was assessed via the one-item Left-Right Self-Placement scale (Breyer, 2015), ranging from 1 = "left" to 10 = "right". [b] Political interest was assessed via a one-item scale (Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V., 2014), ranging from 1 = "not at all" to 10 = "very strongly". [c] Satisfaction with Germany's political system via the German Satisfaction with the Political System Short Scale (SPS; Dentler, Bluemke, & Gabriel, 2020), ranging from 1 = "satisfied" to 4 = "dissatisfied".

S5: Additional Analyses

S5.1: By-Item Analyses Experiment 1

To investigate the generalizability of the inference effect across all statements, we computed a one-tailed dependent samples *t* test for each item independently. Descriptive statistics for implied and implied-other ratings as well as *t* test results can be found in Table S1.1. Inference effects were significant for all 24 items (all *p*s < .001, d_z ranging from 0.78 to 5.36). We take this to mean that participants inferred ideological categories from all the target statements, but to a varying degree.

S5.2: By-Item Analyses Experiment 3

To investigate the generalizability of the discounting effect across all statements, we computed inference effects for each item independently and a two-tailed dependent samples *t* test for each item independently and compared them using two-tailed Welch two-sample *t* tests. Descriptive statistics for the inference effects in the reason and control conditions as well as *t* test results can be found in Table S1.2. Note that the individual discounting effect was significant for only five items. One additional item had a significant but opposite effect (augmentation effect). We

conclude that there was considerable variability in how the items contributed to the discounting effect obtained in Experiment 3.

S5.3: Multiverse Analysis (Experiment 2)

Table S5.3

Multiverse Analysis for the Inference Effect in Experiment 2

		Inference Effect			
Transformation	**Cutoff**	***t***	***df***	***p***	***d_z***
none	none	15.44	161	<.001	1.21
	M + 2 *SD*	16.59	162	<.001	1.30
	2500 ms	18.33	162	<.001	1.44
	2000 ms	18.00	162	<.001	1.41
	1500 ms	20.16	162	<.001	1.58
log	none	20.04	162	<.001	1.57
	M + 2 *SD*	19.03	162	<.001	1.49
	2500 ms	21.49	162	<.001	1.68
	2000 ms	21.06	162	<.001	1.65
	1500 ms	21.92	162	<.001	1.72
inverse	none	21.46	162	<.001	1.68
	M + 2 *SD*	19.00	162	<.001	1.49
	2500 ms	21.91	162	<.001	1.72
	2000 ms	21.50	162	<.001	1.68
	1500 ms	21.26	162	<.001	1.66

Note. Cutoff values indicate the criterion applied to exclude outlier response times prior to analysis (*M* = mean; *SD* = standard deviation). Transformation refers to the mathematical transformation applied to response times prior to analysis (none, log, inverse). *t* = Student's t-statistic; *df* = degrees of freedom; *p* = p-value; d_z = Cohen's d for dependent samples.

S5.4: Error Rates (Experiment 2)

Although we instructed participants to respond as accurately as possible, it is nevertheless possible that the inference effect is also visible in the error rates. We therefore conducted an exploratory error rate analysis. If participants inferred the ideological categories while reading the statements, they should have higher error rates for implied compared to implied other probes. A Wilcoxon signed-rank test showed that participants' error rates were indeed significantly higher in the implied condition (*Mdn* = 0.042) than in the implied other condition (*Mdn* = 0), $Z = 8.38$, $p < .001$, $r = 0.66$.

S5.5: Dislike and Distrust (Experiment 3)

To test whether dislike and distrust were reduced in the reason compared to the control condition, we conducted two paired *t* tests. On average, participants disliked targets in the sufficient condition ($M = 0.089$, $SD = 0.37$) significantly less than in than in the control condition ($M = 0.31$, $SD = 0.33$), $t(199) = -7.54$, $p < .001$, $d_z = 0.53$. Furthermore, they distrusted targets in the sufficient condition ($M = -0.003$, $SD = 0.41$) significantly less than in than in the control condition ($M = 0.18$, $SD = 0.38$), $t(199) = -6.54$, $p < .001$, $d_z = 0.46$.

S5.6: Multiverse Analysis (Experiment 4)

Table S5.6

Multiverse Analysis for the Inference and Discounting Effects in Experiment 4

		Inference Effect					Discounting Effect				
Transformation	**Cutoff**	***DFn***	***DFd***	***F***	***p***	η_p^2	***DFn***	***DFd***	***F***	***p***	η_p^2
none	none	1	232	337.93	<.001	.593	1	232	4.63	.032	.020
	M + 2 *SD*	1	235	340.40	<.001	.592	1	235	2.35	.126	.010
	2500 ms	1	235	437.84	<.001	.651	1	235	0.16	.694	.001
	2000 ms	1	235	485.40	<.001	.674	1	235	0.51	.477	.002
	1500 ms	1	235	512.77	<.001	.686	1	235	0.00	.990	.000
log	none	1	235	537.23	<.001	.696	1	235	0.40	.526	.002
	M + 2 *SD*	1	235	449.71	<.001	.657	1	235	1.44	.232	.006
	2500 ms	1	235	571.26	<.001	.709	1	235	0.05	.828	.000
	2000 ms	1	235	589.61	<.001	.715	1	235	0.19	.662	.001
	1500 ms	1	235	570.20	<.001	.708	1	235	0.03	.874	.000
inverse	none	1	234	655.44	<.001	.737	1	234	0.03	.856	.000
	M + 2 *SD*	1	235	478.79	<.001	.671	1	235	0.91	.342	.004
	2500 ms	1	234	648.57	<.001	.735	1	234	0.02	.897	.000
	2000 ms	1	234	641.82	<.001	.733	1	234	0.00	.981	.000
	1500 ms	1	234	590.36	<.001	.716	1	234	0.20	.651	.001

Note. Cutoff values indicate the criterion applied to exclude outlier response times prior to analysis (*M* = mean; *SD* = standard deviation). Transformation refers to the mathematical transformation applied to response times prior to analysis (none, log, inverse). Results are reported separately for the inference effect and the discounting effect. *DFn* = numerator degrees of freedom; *DFd* = denominator degrees of freedom; *F* = F-statistic; *p* = p-value; η_p^2 = partial eta squared.

S5.7: Error Rates (Experiment 4)

We again conducted an exploratory error rate analysis. In case of discounting the difference between error rates for implied compared to implied-other probes should be smaller for reason compared to control ancillary statements. The main effect for probe type was significant, $F(1, 164) = 191.39$, $p < .001$, $\eta_p^2 = .539$. This again indicates that participants spontaneously inferred the implied labels while reading the statements. However, there was no evidence for discounting, as the interaction between probe type and ancillary statement was not significant, $F(1, 164) = 0.21$, $p = .648$, $\eta_p^2 = .001$.

S5.8: Order of Presentation (Experiment 4)

We conducted a three-way repeated-measures ANOVA to evaluate the effects of probe type, ancillary statement, and order of presentation (behavior first vs. ancillary statement first) on the mean response latencies. Due to a programming error, item order was randomized and not counterbalanced with the other experimental factors. Thus, the observations per cell varied between 0 and 12. There was a significant main effect for probe type, $F(1, 234) = 464.82$, $p < .001$, $\eta_p^2 = .665$, and no interaction between probe type and ancillary statement, $F(1, 234) = 0.50$, $p = .480$, $\eta_p^2 = .002$. There was, however, a marginally significant three-way interaction, $F(1, 234) = 2.91$, $p = .089$, $\eta_p^2 = .012$. Post-hoc analyses revealed that there was a marginally significant interaction between probe type and ancillary statement in the ancillary statement first condition, $F(1, 234) = 2.75$, $p = .099$, $\eta_p^2 = .012$, indicating a trend discounting effect. No such effect was observed when the ancillary statement was presented after the behavioral statement, $F(1, 234) = 0.50$, $p = .481$, $\eta_p^2 = .002$.

S5.9: Stimulus Set Version (Experiment 5)

For Experiment 5 we created two parallel versions of the stimulus set. In version A the first half of the 24 target faces were paired with implied labels and the second half with implied-other labels, whereas in version B it was the other way around. There were no theoretically meaningful differences between both stimulus set versions. We nevertheless found effects of the stimulus set

version: a main effect, $F(1, 384) = 8.40$, $p = .004$, $\eta_p^2 = .021$, a two-way interaction with probe type, $F(1, 384) = 17.99$, $p < .001$, $\eta_p^2 = .045$, and a three-way interaction with probe type and ancillary statement , $F(1, 384) = 5.47$, $p = .020$, $\eta_p^2 = .014$. Post-hoc tests revealed that the main effect of probe type was significant for both stimulus set version A, $F(1, 195) = 61.19$, $p < .001$, $\eta_p^2 = .239$, and stimulus set version B, $F(1, 189) = 169.54$, $p < .001$, $\eta_p^2 = .473$. This suggests that participants inferred ideological categories across both stimulus set versions. The interaction between probe type and ancillary statement was not significant for stimulus set version A, $F(1, 195) = 0.52$, $p = .472$, $\eta_p^2 = .003$, but did reach significance in stimulus set version B, $F(1, 189) = 6.86$, $p = .010$, $\eta_p^2 = .035$, indicating a significant discounting effect in set B. The fact that there was evidence of discounting only for one stimulus set version seems to suggest, that discounting occurred for only a subset of items. However, because each item's implied and implied-other trials were distributed across the two stimulus set versions, these results cannot be directly interpreted in terms of individual item-effects.

References

Breyer, B. (2015). Left-right self-placement scale (ALLBUS/GGSS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis83

Dentler, K., Bluemke, M., & Gabriel, O. W. (2020). German Satisfaction with the Political System Short Scale (SPS). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/zis278

Zentralarchiv für empirische Sozialforschung (ZA) & Zentrum für Umfragen, Methoden und Analysen (ZUMA) e.V. (1998). Politikinteresse. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/zis181

SUPPLEMENTARY MATERIALS II

Second Article

Table of Contents

S1: Pilot Study **179**

Methods 179

Results 182

Discussion 183

S2: Categorization Items **186**

S3: Example Transcript **187**

S4: Additional Analyses of Perceived Prototypicality **191**

S5: Manipulation Checks **191**

Multiverse Analysis for Experiment 2 192

Multiverse Analysis for Experiment 3 193

S6: Experiment 1 Analyses with Outlier Exclusion **194**

Manipulation Check 194

Party Categorization 194

S7: Experiment 3 Analyses Without Outlier Exclusion **196**

S8: Moderation by Partisanship **196**

S9: Expected Effect Size for Stereotype Probability Condition **197**

References **198**

S1: Pilot Study

In the pilot study we presented a US-American sample with a purportedly real newspaper article about a local hero who belongs to one of two party categories—Republicans or Democrats. This category exemplar expressed either a stereotype-confirming or stereotype-disconfirming opinion about a highly polarized issue. Analyses were performed using the ordinal package (Christensen, 2023).

Methods

Sample

The sample size was determined using a simulation, based on pilot data showing a 56% reduction for the odds of categorizing the target in the stereotype-disconfirming versus stereotype-confirming condition. With $\alpha = .05$ and a targeted power of 95%, the minimum sample size needed with this effect size was $N = 384$ for an ordinal regression. We therefore collected a sample of $N = 384$ valid datasets (gender: 161 women, 216 men, 3 non-binary, 4 undisclosed; age: *Mdn* = 39, ranging from 19 – 80; partisanship: 202 Democrats, 90 Republicans, 92 Independents).

Procedure

We told participants that they would be presented with an excerpt of a newspaper article taken from one of several newspapers. We asked them to read the article and then answer some questions about it. In the article the target person identified with a political party (i.e., the target person stated to be a registered Republican or Democrat) and talked about their political views. Depending on the typicality condition the target either expressed a stereotype-confirming view or a stereotype-disconfirming view on a polarized political target issue (e.g., abortion). In the neutral condition no view regarding the target issue was expressed nor was the target issue mentioned.

After reading the text, participants indicated their agreement to 11 items. All items were presented on a single page with a 7-point scale ranging from 1 = “strongly disagree” to 7 = “strongly agree”. The first four items were fillers, followed by two items measuring political categorization. The first categorization item pertained to the opinion that is common in the target person’s

category. For instance, if the target person was presented as a Republican talking about abortion, the first item would state “People who are pro-life are extremely likely to vote Republican”. This item served to measure the effect of reducing the stereotype probability (i.e., reducing the probability of a Republican being pro-life). The second categorization item pertained to the opinion that is uncommon in the target person’s category. Thus, in our example it would state “People who are pro-choice are extremely likely to vote Democrat”. This item was our main dependent measure and served to measure the effect of increasing the counterstereotype probability (i.e., increasing the probability of a Republican being pro-choice). This item is presented second, because it features a previously unmentioned category (i.e., Democrats), which might confuse participants if it was presented first.

The categorization items were followed by another filler item, two items measuring structural differentiation (Koudenburg & Kashima, 2022) which were separated by yet another filler item, and an item measuring stereotypes of the target person’s category (e.g., “On average, Republicans strongly believe that abortion should be illegal.”). The stereotype item served as the manipulation check. After the block of items, participants were asked about their perceptions of ideological polarization. Specifically, we included two items from Ploger (2024) that assessed perceived public ideological polarization (“How divided do you think most ordinary Americans are on political issues facing the country?”) and perceived partisan ideological polarization (“How divided do you think most Democrat and Republican voters are on political issues facing the country?”).

Next, we asked participants for demographic data, including their gender, age, and political identity (Democrat, Republican, Independent), whether they thought their data was suitable for analysis (yes or no), whether anything about the newspaper article stood out to them, and whether they have an idea about the specific aim of the study. Ultimately, they were debriefed about the purpose of the study and asked for their informed consent for data storage and analysis. The whole experiment lasted approximately 5 minutes.

Materials

The presented article tells the story of a local hero who saved a family from a burning house. The first half of the article describes the target person's heroic actions. The target person then shares his thoughts about the events in an interview and criticizes the city council for not ensuring adequate funding of the fire department, which prompts the interviewer to ask about his political engagement. In all three typicality conditions the target person responds by saying that he was a registered Republican [Democrat] but was not deeply engaged in politics. In the neutral condition the article ends with some information about his personal life. In both the stereotype-confirming and stereotype-disconfirming condition, the article states that the target person raises some issues that interested him and expressed the typical opinions of his party. This was implemented to prevent subtyping by ensuring that the target person is perceived as a typical party member. Next, the article states that the issue of [target issue] seemed particularly important to the target person. The target person then expresses the stereotype-confirming or stereotype-disconfirming opinion (e.g., "Personally, I think abortion should be illegal [legal]") and then gives a relatively vague explanation for his opinion (e.g., "I've had experiences that made me realize the importance of protecting unborn life [having that choice]"). Lastly, he says that he knows "quite a few fellow Republicans [Democrats] who share a similar perspective, but, of course, there's also people who lean the other way". This sentence was added to serve as a meta-cognitive trigger for cognitive inclusion of the exemplar into the category (McIntyre et al., 2016). The article then ends with some information about his personal life as in the neutral condition.

The first part of the article (i.e., the local hero story) was identical to all participants. The second part differed between participants. Specifically, we varied the target issues (abortion, gun control, income equality, immigration, welfare, gay marriage, environmental protection, affirmative action), target categories (Republican, Democrat), and the typicality of the opinion (stereotype-confirming, neutral, stereotype-disconfirming), yielding a total of 8 * 2 * 3 = 48 articles. The target issues were taken from Lüders et al. (2023) who showed these issues to be highly polarized in the US

context. A spreadsheet with the 48 different versions of article can be found in this paper's osf-project. Category labels and opinions in the test items were varied to match the contents of the article.

Results

For our manipulation check, we fitted a cumulative link mixed model with the Laplace approximation. We included the stereotype item as the dependent variable, fixed effects for typicality and target partisanship, and a random effect for issue. Due to a programming error, the wrong stereotype items were displayed for the issue of environmental protection. We therefore excluded this issue from the analysis. The stereotype in the stereotype-disconfirming condition was not statistically weaker than in the neutral condition (z = 1.48, p = .139). The contrast between the stereotype-disconfirming and stereotype-confirming condition however was statistically significant (z = 2.20, p = .028). An additional contrast between the neutral and stereotype-confirming condition was not significant (z = 0.74, p = .460). There was no effect of target partisanship (z = 0.053, p = .958). Although these significant contrast between the stereotype-disconfirming and stereotype-confirming condition suggests that our manipulation was successful, the fact that the contrasts with the neutral condition remained non-significant indicates that the effect of the manipulation was rather small.

For our preregistered analysis we fitted another cumulative link mixed model with the Laplace approximation. We included the categorization item as the dependent variable, fixed effects for typicality and target partisanship, and a random effect for issue.[1] Contrary to our hypothesis, political categorization in the stereotype-disconfirming condition was neither significantly weaker than in the neutral condition (z = 1.15, p = .252) nor than in the stereotype-confirming condition (z = 0.82, p = .411). There was no effect of target partisanship (z = −0.84, p = .400).

[1] We had accidentally preregistered target partisanship as a random effect, but as this variable has only two levels, we instead included it as a fixed effect.

To explore the effect of reducing stereotype probability we fitted another cumulative link mixed model with the Laplace approximation. We included the other categorization item (measuring the effect of changing the stereotype probability) as the dependent variable, fixed effects for typicality and target partisanship, and a random effect for issue. Political categorization in the stereotype-disconfirming condition was both significantly weaker than in the neutral condition ($z = 2.61$, $p = .009$) and in the stereotype-confirming condition ($z = 3.38$, $p < .001$). An additional contrast between the neutral and stereotype-confirming condition was not significant ($z = -0.74$, $p = .459$. There was no effect of target partisanship ($z = -0.82$, $p = .413$).

Discussion

Contrary to our expectations, we found no significant effects of typicality on political categorization in our main analysis. However, we did find a significant effect of typicality in the exploratory analysis investigating the effect of reducing the stereotype probability. This is surprising, because the expected effect of reducing the stereotype probability is smaller than that of increasing the counterstereotype probability. One potential explanation of this puzzling finding may be that participants were confused by the relatively abstract phrasing of the categorization items. As a reminder, the items were not phrased as open questions but as propositions that participants with reduced political categorization would have to negate. The high cognitive demand of the logical analysis afforded by these items may have been met with the use of mental shortcuts, i.e., with participants concluding that because they perceived information solely about members of the target category, they could not apply this information to an item about the opposing category. Items that ask participants how likely it is that someone given a certain opinion on the target issue belongs to either of the two categories may be easier to understand and therefore better suited to assess participants' intuitive categorizations.

Moreover, the rather small effect in the manipulation check suggests that the typicality manipulation was not very strong. One possible explanation for its weak effect is that participants may have perceived the article to be fake or even suspected that it was supposed to manipulate

their beliefs. In line with this idea, responses to open-ended questions from a funnel debriefing indicate that many participants found it unusual that a human-interest story about a local hero also talked about politics. Doubting the article's authenticity might have prevented participants from including the exemplars in their category representations. Thus, it may be preferable to present the exemplar in a context where the expression of partisanship or political opinions is less surprising.

Moreover, it may be that the target person was perceived as atypical. Whether a counterstereotype leads to stereotype change crucially depends on whether the counterstereotypical exemplar is included in the category or excluded from it. Cognitive exclusion may occur when an exemplar is perceived to be atypical for a category (subtyping) and might even lead to stronger stereotypes via contrast effects (Bless & Schwarz, 1998; Kunda & Oleson, 1997).

Although we described the exemplar as being generally in accordance with the category views, he may nevertheless have been perceived as atypical because he was described as a local hero, which already implies that he is in some sense exceptional. Research has shown that even attributes that are neutral with respect to the examined category can facilitate subtyping (Kunda & Oleson, 1995). Thus, the stereotype-reducing effect of the intervention could potentially be enhanced by ridding the cover story of superfluous individuating information. Whether the exemplar is perceived as atypical may also depend on prior expectations (Kunda & Oleson, 1997). A person with very extreme party stereotypes may be more likely to perceive a counterstereotypical exemplar as atypical than a person with moderate party stereotypes. Thus, the same manipulation could reduce stereotypes of some participants while increasing those of others.

In Experiment 1 we address the limitations of the pilot study in the following ways: First, we present the category exemplar not via a newspaper article but via a transcript of a purportedly real radio show in which two ordinary people discuss questions about politics. In this context the expression of partisanship or political opinions should be less surprising compared to a human-interest story, making the intervention more believable. Second, we use a more straightforward categorization measure, in which we present participants with a certain opinion and ask them how

likely a person with this opinion is to belong to either category. Third, we use a pre-post design, measuring political categorization before and after the manipulation. In case the manipulation fails, this allows us to check whether there are opposing effects for individuals with different prior levels of stereotypes. For the sake of brevity, we only assess political categorization based on the opinion that is uncommon for the target person's category, i.e., we only examine the effect of increasing the counterstereotype probability.

S2: Categorization Items

Table S2

Items Used in Experiments 1, 2, and 3

Item Type	Issue	Original Item	Reversed Item
test	Gun control	The federal government should make it more difficult to buy a gun.	The federal government should protect and uphold Second Amendment rights
test	Environmental protection	The government should regulate business to protect the environment.	The government should eliminate environmental regulations and allow businesses to self-regulate.
test	Affirmative action	The federal government should make a concerted effort to improve social and economic conditions for African Americans.	The federal government should promote equal opportunity for all Americans instead of giving special treatment to some racial groups.
filler	Income equality	The government should take steps to make incomes more equal.	The government should reduce regulations to let the free market set incomes.
filler	Immigration	All unauthorized immigrants should be sent back to their home country.*	Unauthorized immigrants should be provided a pathway to citizenship.
filler	Abortion	Abortion should be illegal.*	Abortion should be legal.

Note. Original items were taken from Lüders et al. (2023) which were phrased in a way such that agreement with an item is generally diagnostic of being a Democrat except for the two items marked with an asterisk, for which agreement is diagnostic of being a Republican. We thus created reversed versions of these items. The diagnosticity of the test items was manipulated across participants (i.e., for each participant the diagnosticity of one of the test items was manipulated). The diagnosticity of the filler items was not manipulated, but the respective issues were mentioned in the transcript.

S3: Example Transcript

Figure S3.1

Example Radio Show Transcript with Highlighted Typicality Manipulation



Round 2

Now you will be given some statements and asked to discuss your opinions about them. The first one is: "The government should reduce regulations to let the free market set incomes." — Do you agree or disagree?

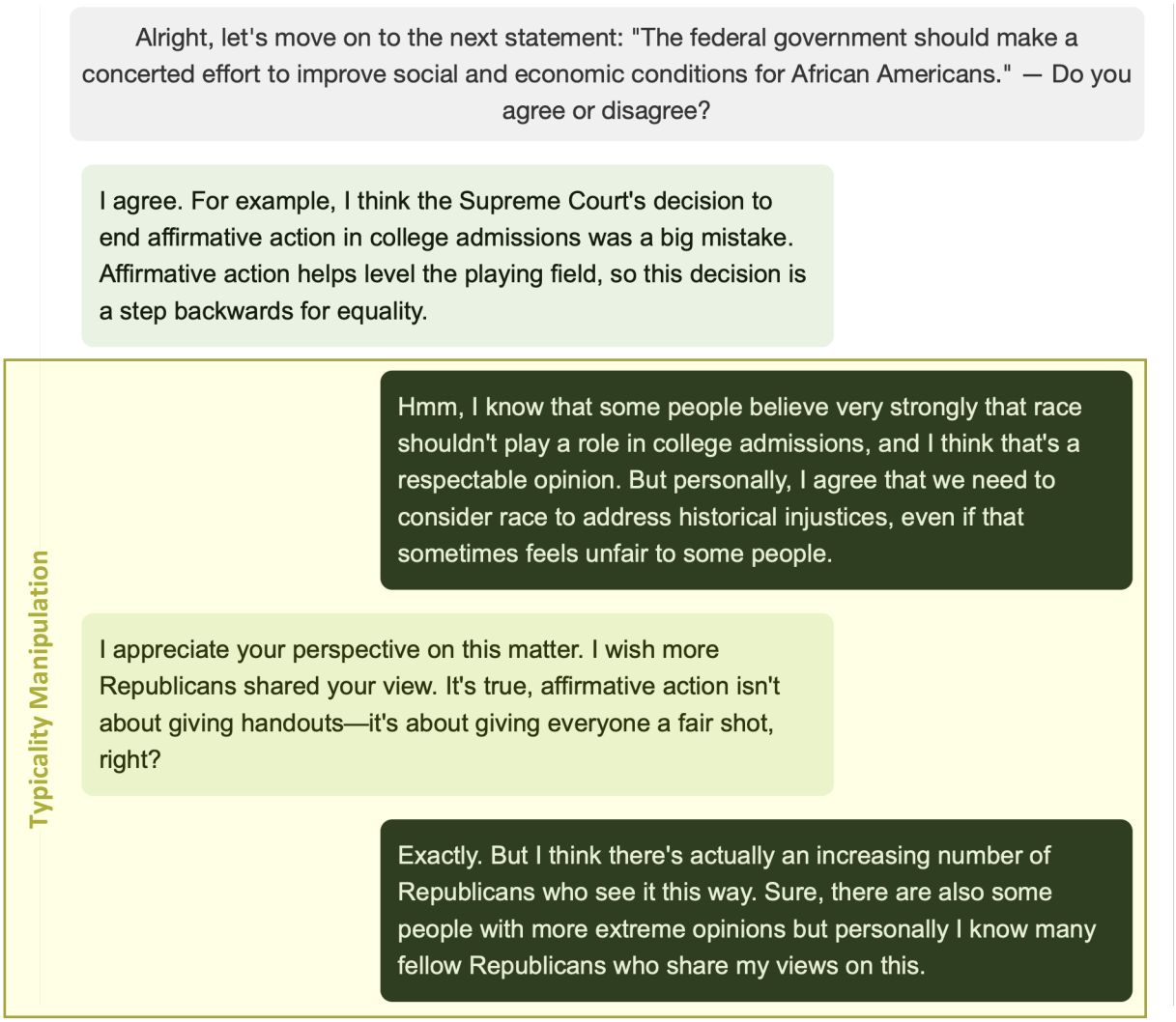
I disagree with the statement. I think income inequality is a big problem. It's not fair that some people earn more money than they could ever spend while others don't even have enough to buy food for their children.

I agree with you that poverty is a problem, but I nevertheless agree with the statement. I think it's reasonable that people with a better education and greater responsibility earn more money. The solution in my view is that we need to grow our economy and bring back industrial jobs from abroad.

I mean sure, there's nothing wrong with some people earning a bit more when they are highly educated but it can't be that CEOs earn 300 times more than the average worker. And this problem won't be fixed by bringing back some industrial jobs.

I don't see how high salaries of CEOs are a problem. Companies need to recruit the most qualified people and they do that by making the best offers. The problem isn't that CEOs earn too much. It's that many people earn too little.

I think these two things are not really separable. If you want to increase salaries for the average worker, you need to take away the money somewhere.



Note. This is an example transcript which was presented to participants in the stereotype-disconfirming condition, with the target issue being affirmative action and the target category being Republican. Typicality was manipulated within the last three responses (highlighted). The corresponding responses in the stereotype-confirming condition can be found in Figure S3.2.

Figure S3.2

Typicality Manipulation in the Stereotype-Confirming Condition



Note. This figure shows the stereotype-confirming version of the last three responses in Figure S3.1 (i.e., of the typicality manipulation). The target issue is again affirmative action and the target category being Republican. The rest of the transcript was identical across typicality conditions.

S4: Additional Analyses of Perceived Prototypicality

We analyzed the perceived prototypicality ratings for Experiments 2 and 3 together. The target person's perceived prototypicality in the stereotype-disconfirming condition (M = 57.1, SD = 22.6) was significantly lower than in the stereotype-confirming condition (M = 81.5, SD = 16.2), $t(837.49) = -18.9$, $p < .001$, $d = -1.24$, 90% CI [−1.36, −1.12]. To explore the effect of perceived prototypicality on political categorization, we ran a linear regression model with political categorization as the dependent variable (again excluding outliers with categorization scores below 40) and typicality, perceived prototypicality, and the interaction between the two as predictors. We used treatment-coding for typicality and scale midpoint centering for perceived prototypicality. The overall model was significant, $F(3, 921) = 37.0$, $p < .001$, $R^2 = .11$. The main effect of typicality was not significant, $b = -0.18$, $SE = 1.59$, $t(921) = -0.11$, $p = .912$, indicating that political categorization was not weaker in the stereotype-disconfirming condition at medium perceived prototypicality. The simple main effect of perceived prototypicality was significant, $b = 0.24$, $SE = 0.04$, $t(921) = 5.92$, $p < .001$, indicating that in the stereotype-confirming condition higher perceived prototypicality was associated with stronger political categorization. Importantly, the interaction effect was significant, $b = -0.28$, $SE = 0.05$, $t(921) = -5.58$, $p < .001$, indicating that in the stereotype-disconfirming condition this relation was canceled out.

S5: Manipulation Checks

We report additional analyses of the manipulation checks in Experiments 2 and 3 investigating the effects of outlier exclusion on the results. As there is no obvious cutoff criterion, we conducted a multiverse analysis using several different cutoff criteria. We excluded participants with values below 1, 10, 20, 30, or 40 (on a scale ranging from 0 to 100) and conducted two-way between-subjects ANOVAs with the stereotype measure as the dependent variable and typicality and category type as between-subjects factors for each of the cutoff criteria. For the results of the multiverse analysis see Table S5.1 (Experiment 2) and Table S5.2 (Experiment 3).

Multiverse Analysis for Experiment 2

Table S5.1

Results of a Multiverse Analysis for the Manipulation Check in Experiment 2

				Category Type			Typicality			Interaction		
cutoff	$n_{excluded}$	***DFn***	***DFd***	***F***	***p***	η_p^2	***F***	***p***	η_p^2	***F***	***p***	η_p^2
none	0	1	356	0.01	.918	.000 [.000, .003]	3.32	.069	.009 [.000, .032]	0.58	.447	.002 [.000, .016]
fixed 1	13	1	343	0.06	.800	.000 [.000, .008]	7.62	**.006**	.022 [.004, .053]	2.41	.121	.007 [.000, .029]
fixed 10	19	1	337	0.00	.996	.000 [.000, .000]	7.01	**.009**	.020 [.003, .051]	1.89	.171	.006 [.000, .026]
fixed 20	33	1	323	0.13	.718	.000 [.000, .011]	10.81	**.001**	.032 [.008, .070]	0.03	.871	.000 [.000, .006]
fixed 30	59	1	297	1.43	.233	.005 [.000, .026]	6.71	**.010**	.022 [.003, .056]	0.32	.570	.001 [.000, .016]
fixed 40	70	1	286	0.29	.591	.001 [.000, .016]	5.85	**.016**	.020 [.002, .054]	0.52	.471	.002 [.000, .019]

Note. We conducted two-way between-subjects ANOVAs assessing the effects of typicality and category type on stereotyping across different cutoff criteria. Columns represent the number of excluded participants ($n_{excluded}$), degrees of freedom (*DFn*; *DFd*), *F* statistics, *p* values, and partial eta squared (η_p^2) for each effect. Significant results are highlighted ($p < .05$).

Multiverse Analysis for Experiment 3

Table S5.2

Results of a Multiverse Analysis for the Manipulation Check in Experiment 3

				Category Type			Typicality			Interaction		
cutoff	$n_{excluded}$	***DFn***	***DFd***	***F***	***p***	η_p^2	***F***	***p***	η_p^2	***F***	***p***	η_p^2
none	0	1	628	1.75	.186	.003 [.000, .014]	17.19	**<.001**	.027 [.010, .051]	1.22	.269	.002 [.000, .012]
fixed 1	6	1	622	0.86	.353	.001 [.000, .010]	18.54	**<.001**	.029 [.011, .054]	0.89	.345	.001 [.000, .011]
fixed 10	12	1	616	0.56	.455	.001 [.000, .009]	21.86	**<.001**	.034 [.014, .061]	0.62	.431	.001 [.000, .009]
fixed 20	33	1	595	0.26	.610	.000 [.000, .007]	33.04	**<.001**	.053 [.027, .084]	0.62	.430	.001 [.000, .010]
fixed 30	67	1	561	0.01	.912	.000 [.000, .002]	48.45	**<.001**	.079 [.047, .117]	0.57	.449	.001 [.000, .010]
fixed 40	100	1	528	0.36	.549	.001 [.000, .009]	45.11	**<.001**	.079 [.045, .117]	0.79	.374	.001 [.000, .012]

Note. We conducted two-way between-subjects ANOVAs assessing the effects of typicality and category type on stereotyping across different cutoff criteria. Columns represent the number of excluded participants ($n_{excluded}$), degrees of freedom (*DFn*; *DFd*), *F* statistics, *p* values, and partial eta squared (η_p^2) for each effect. Significant results are highlighted ($p < .05$).

S6: Experiment 1 Analyses with Outlier Exclusion

Manipulation Check

To test whether the results of Experiment 1 change when applying the same outlier cutoff as in Experiments 2 and 3, we ran an exploratory multiverse analysis with multiple fixed outlier cutoffs. We excluded participants with values below 1, 10, 20, 30, or 40 and conducted a one-sided independent samples *t* test comparing stereotyping in the stereotype-confirming and stereotype-disconfirming conditions. Across all outlier thresholds, stereotyping was significantly weaker when stereotypes were disconfirmed as opposed to when they were confirmed.

Table S6

Results of a Multiverse Analysis for the Manipulation Check in Experiment 1

cutoff	$n_{excluded}$	*df*	*t*	*p*	*d* [CI 90%]
none	0	212.05	−3.08	**.001**	−0.42 [−0.67, −0.20]
fixed 1	4	209.93	−3.83	**<.001**	−0.53 [−0.79, −0.29]
fixed 10	6	207.06	−4.48	**<.001**	−0.62 [−0.88, −0.39]
fixed 20	8	204.30	−4.64	**<.001**	−0.64 [−0.90, −0.42]
fixed 30	17	193.07	−4.70	**<.001**	−0.67 [−0.91, −0.42]
fixed 40	28	181.38	−4.07	**<.001**	−0.59 [−0.87, −0.36]

Note. We conducted two-way between-subjects ANOVAs assessing the effects of typicality and category type on stereotyping across different cutoff criteria. Columns represent the number of excluded participants ($n_{excluded}$), degrees of freedom (*df*), *t*-values, *p*-values, and Cohen's *d* with 90% confidence intervals in brackets. Significant results are highlighted (p < .05).

Party Categorization

To test whether the results change when applying the same outlier cutoff as in Experiments 2 and 3, we ran an exploratory two-way mixed ANOVA with political categorization as the dependent variable, time as a within-subjects factor, and typicality as a between-subjects factor. After excluding 19 participants with outlying values, we obtained the same pattern of results as without outlier exclusion, namely a significant effect of typicality, $F(1, 195) = 12.12$, $p < .001$, $\eta_p^2 = .059$, 90% CI [.016,

.119], and time, $F(1, 195) = 5.50$, $p = .020$, $\eta_p^2 = .027$, 90% CI [.002, .075], qualified by a significant interaction, $F(1, 195) = 44.93$, $p < .001$, $\eta_p^2 = .187$, 90% CI [.111, .265]. Simple effects analyses again revealed that the effect of time was significant and negative in the stereotype-disconfirming condition, $t(94) = -5.45$, $p < .001$, $d_z = -0.56$, 95% CI [−0.75, −0.40], indicating that the counterstereotypical exemplars did reduce political categorization compared to the baseline level. In the stereotype-confirming condition the effect was significant and positive, $t(101) = 3.78$, $p < .001$, $d_z = 0.37$, 95% CI [0.18, 0.58], indicating that stereotypical exemplars increased political categorization compared to the baseline level. However, the effect size in the stereotype-disconfirming condition was significantly larger than in the stereotype-confirming condition, $t(169.77) = 2.32$, $p = .022$, $d = 0.33$, 95% CI [0.060, 0.62], suggesting the interaction was more strongly driven by the decrease in political categorization due to the counterstereotypical exemplars than by the increase due to the stereotypical ones. Effect sizes for the interaction and simple effects were generally larger than in the analysis without outlier exclusion.

We also repeated the analyses for the non-target items with the outlier cutoff. As in the main analysis, we took the average of the two test items for which diagnosticity had not been manipulated for the respective participant and used the resulting score as the dependent variable of a two-way mixed ANOVA with time as a within-subjects factor and typicality as a between-subjects factor. After excluding 4 participants with outlying values, we obtained the same pattern as without outlier exclusion, that is, a small significant effect of typicality, $F(1, 210) = 4.10$, $p = .044$, $\eta_p^2 = .019$, 90% CI [.000, .060], a non-significant effect of time, $F(1, 210) = 0.91$, $p = .340$, $\eta_p^2 = .004$, 90% CI [.000, .031], and a non-significant interaction, $F(1, 210) = 0.38$, $p = .540$, $\eta_p^2 = .002$, 90% CI [.000, .023].

S7: Experiment 3 Analyses Without Outlier Exclusion

We re-ran the main analyses without the exclusion of values below 40. In the counterstereotype condition, the effect of typicality remained significant for both party categories, $t(314.62) = -3.13$, $p < .001$, $d = -0.35$, 90% CI [−0.56, −0.17], and ideological categories, $t(312.24) = -4.16$, $p < .001$, $d = -0.47$, 90% CI [−0.68, −0.27]. However, both effects got notably smaller when including outliers. In the stereotype probability condition the effect of typicality was no longer significant for party categories, $t(314.62) = -1.11$, $p = .135$, $d = -0.12$, 90% CI [−0.31, 0.060], but remained significant for ideological categories, $t(305.83) = -3.59$, $p < .001$, $d = -0.40$, 90% CI [−0.60, −0.23].

S8: Moderation by Partisanship

To test whether the effect of typicality significantly differed between Republican, Democratic, and Independent participants, we conducted a two-way between-subjects ANOVA with political categorization as the dependent variable, and typicality and participants' party identification as between-subjects factors. For this analysis we combined the data from Experiments 2 and 3. We obtained a significant medium effect of typicality, $F(1, 890) = 67.42$, $p < .001$, $\eta_p^2 = .070$, 90% CI [.045, .098] and a small significant effect for party identification, $F(2, 890) = 4.55$, $p = .011$, $\eta_p^2 = .010$, 90% CI [.001, .022]. Importantly, these effects were not qualified by an interaction, $F(2, 890) = 0.14$, $p = .869$, $\eta_p^2 = .000$, 90% CI [.000, .002], suggesting that our manipulation was comparably effective for Republican, Democratic, and Independent participants.

S9: Expected Effect Size for Stereotype Probability Condition

Based on the diagnosticity formula, manipulating the stereotype probability should have a smaller effect on political categorization than manipulating the counterstereotype probability—at least for highly diagnostic cues. To quantify this difference, we ran a simulation estimating the expected effects of manipulating the stereotype probability given the observed effects of manipulating the counterstereotype probability. That is, we calculated the expected condition means for a range of different baseline probabilities $\boldsymbol{P}(\boldsymbol{cue\ C}|\boldsymbol{category\ A})$ and $\boldsymbol{P}(\boldsymbol{cue\ C}|\boldsymbol{category\ B})$ and intervention effects (i.e., changes in these probabilities)—assuming that a perceived diagnosticity of 0.5 corresponds to a categorization score of 50 and a perceived diagnosticity of 1 to a categorization score of 100. We then simulated 10,000 datasets based on the observed means and variances in the counterstereotype probability condition. To account for the skewness of the data, we used a beta distribution. For each simulated dataset we selected the combination of baseline probabilities and intervention effects which produced the expected condition means closest to the simulated condition means. We then averaged the selected baseline probabilities, intervention effects, and expected effects for the counterstereotype and stereotype probability conditions. This procedure suggested that for a population with the observed means and variances the stereotype probability is 79%, the counterstereotype probability is 12%, and the change in probabilities is 9%. Given these values the expected difference in the stereotype probability conditions (–1.6) should be 5 times smaller than the difference in the counterstereotype probability condition (–8.0). To test whether the effect observed in the stereotype probability condition differed from the expected effect, we conducted a two-sided independent samples *t* test, comparing the stereotype-confirming and stereotype-disconfirming conditions, but with the null hypothesis that the observed difference (–4.5) equals the expected difference (–1.6) rather than zero.

References

Bless, H., & Schwarz, N. (1998). Context effects in political judgement: Assimilation and contrast as a function of categorization processes. *European Journal of Social Psychology*, *28*(2), 159–172. https://doi.org/10.1002/(SICI)1099-0992(199803/04)28:2<159::AID-EJSP860>3.0.CO;2-4

Christensen, R. H. B. (2023). *ordinal: Regression models for ordinal data* (Version 2022.11-16) [Computer software]. https://CRAN.R-project.org/package=ordinal

Koudenburg, N., & Kashima, Y. (2022). A polarized discourse: Effects of opinion differentiation and structural differentiation on communication. *Personality and Social Psychology Bulletin*, *48*(7), 1068–1086. https://doi.org/10.1177/01461672211030816

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*(4), 565–579. https://doi.org/10.1037/0022-3514.68.4.565

Kunda, Z., & Oleson, K. C. (1997). When exceptions prove the rule: How extremity of deviance determines the impact of deviant examples on stereotypes. *Journal of Personality and Social Psychology*, *72*(5), 965–979. https://doi.org/10.1037/0022-3514.72.5.965

Lüders, A., Carpentras, D., & Quayle, M. (2023). Attitude networks as intergroup realities: Using network-modelling to research attitude-identity relationships in polarized political contexts. *British Journal of Social Psychology*, *63*, 37–51. https://doi.org/10.1111/bjso.12665

McIntyre, K., Paolini, S., & Hewstone, M. (2016). Changing people's views of outgroups through individual-to-group generalisation: Meta-analytic reviews and theoretical considerations. *European Review of Social Psychology*, *27*(1), 63–115. https://doi.org/10.1080/10463283.2016.1201893

Ploger, G. (2024). Polarization all the way down: How coverage of elite and partisan polarization spills over to perceptions of the U.S. mass public. *Political Communication*, *41*(3), 393–412. https://doi.org/10.1080/10584609.2024.2312867

SUPPLEMENTARY MATERIALS III

Third Article

Table of Contents

S1: Data Collection and Preprocessing 203

S2: Description of the Training Data 204

S3: Confusion Matrices 205

Label Target 205

Linguistic Form 205

Generics 206

Partisan Prejudice in the Posts 206

Partisan Prejudice in the Comments 207

S4: Interdependence of Partisan Prejudice and Genericity 207

S5: Correlations between Predictors 209

Zero-Order Correlations 209

Between-Surface Correlations 210

Within-Surface Correlations 210

S6: Angry Reactions 211

S7: Top Ten News Media Outlets 215

S8: Moderation Analyses 216

S9: Selection Effects 220

S10: Ingroup Versus Outgroup Labels 223

References 224

S1: Data Collection and Preprocessing

The publication timeframe from November 1, 2023 to October 31, 2024 was selected to balance sample size with three key constraints: MCL API rate limits, computational demands of transformer-based text analysis, and project timelines. Since our classifiers could only categorize entire posts rather than individual label uses, each post needed to address just one category to avoid ambiguity. We therefore retrieved only posts containing exactly one of the following labels (case-insensitive): *Republicans*, *Republican*, *GOP*, *Democrats*, *Dems*, *Democrat*, or *Democratic*. We selected these terms because they are commonly used and relatively specific to their intended political categories, unlike terms such as *Reps*, which could refer to representatives or repetitions rather than Republicans. To ensure contextual relevance, we retrieved only English-language posts published on U.S. *surfaces* (Facebook pages or profiles). To obtain sufficient data points for multilevel analysis, we targeted posts with at least 10 comments. Since the API does not allow pre-retrieval filtering by comment count, we limited our search to posts with a minimum of 5,000 views, which should be relatively likely to have sufficient comments.

In total, we retrieved N = 120,602 posts that fulfilled the described pre-selection criteria. We applied minimal preprocessing to the post texts, removing URLs and converting special Unicode characters (such as 𝒎𝒂𝒕𝒉𝒆𝒎𝒂𝒕𝒊𝒄𝒂𝒍 𝒔𝒄𝒓𝒊𝒑𝒕 𝒍𝒆𝒕𝒕𝒆𝒓𝒔 or **bold letters**) back into plain ASCII letters. We then excluded posts with multiple occurrences of the same label (to avoid classification ambiguity), posts without label occurrences after preprocessing (removing posts where the label occurred in the URL), posts exceeding 200 words (to reduce computational demands), posts with less than 10 comments, and duplicate posts (retaining those with the highest comment count). We then removed posts where labels did not refer to people (excluding the else category from the label target classifier).

For each remaining post, we retrieved up to 20 top-level comments, starting with the oldest. After applying the same preprocessing as for the posts, we removed comments exceeding 200 words, non-English comments, comments with no text, and multiple comments by the same user

(retaining only the first). To ensure sufficient comments per post after these exclusions, we once more removed all posts with fewer than 10 comments, along with any associated comments. This resulted in a final sample of *N* = 45,780 posts and *N* = 769,684 comments.

S2: Description of the Training Data

The posts and comments used for classifier training were retrieved via the Meta Content Library (MCL) and drawn from a distribution similar to the target sample, with two notable differences. First, training data spans April–May 2024, whereas the target sample covers November 2023–October 2024. This shortened timeframe reflects practical constraints, including rate limit restrictions. While we do not expect substantial changes in how variables are expressed over this one-year period, classifier accuracy may decline for data considerably outside the training window. Second, post-level classifiers deviate from the target sample in another regard: Due to delayed API access, training posts were downloaded via the Meta Content Library (MCL) interface rather than the MCL API, limiting us to the "downloadable subset" of posts from Pages with at least 15,000 likes or followers. This restriction does not apply to our main analyses, which use API data. Otherwise, training posts and comments were subject to the same inclusion criteria, exclusion criteria, and preprocessing as the target sample. Datasets were hand-coded by the first author and then split into a training set (80%), a validation set for comparing model performance for different training routines (10%), and a test set for final model evaluation (10%), using stratified sampling to preserve class balance.

S3: Confusion Matrices

Label Target

Table S3.1

Confusion Matrix for Label Target Classifier

	Predicted				
Actual	Individual	Subgroup	Group	Else	Total
Individual	94	1	1	4	100
Subgroup	2	74	2	1	79
Group	1	4	73	1	79
Else	2	4	2	134	142
Total	99	83	78	140	400

Note. This confusion matrix shows prediction accuracy of the Label Target classifier. Rows represent actual class labels; columns represent predicted class labels. Diagonal values indicate correct predictions, while off-diagonal values show misclassifications.

Linguistic Form

Table S3.2

Confusion Matrix for Linguistic Form Classifier

	Predicted		
Actual	Adjective	Noun	Total
Adjective	111	0	111
Noun	1	160	160
Total	112	160	171

Note. This confusion matrix shows prediction accuracy of the linguistic form classifier. Rows represent actual class labels; columns represent predicted class labels. Diagonal values indicate correct predictions, while off-diagonal values show misclassifications.

Generics

Table S3.3

Confusion Matrix for Genericity Classifier

	Predicted		
Actual	Non-Generic	Generic	Total
Non-Generic	198	15	213
Generic	14	73	87
Total	212	88	300

Note. This confusion matrix shows prediction accuracy of the genericity classifier. Rows represent actual class labels; columns represent predicted class labels. Diagonal values indicate correct predictions, while off-diagonal values show misclassifications.

Partisan Prejudice in the Posts

Table S3.4

Confusion Matrix for Partisan Prejudice (Posts) Classifier

	Predicted		
Actual	Non-Prejudiced	Prejudiced	Total
Non-Prejudiced	120	17	137
Prejudiced	8	55	63
Total	128	72	200

Note. This confusion matrix shows prediction accuracy of the partisan prejudice (posts) classifier. Rows represent actual class labels; columns represent predicted class labels. Diagonal values indicate correct predictions, while off-diagonal values show misclassifications.

Partisan Prejudice in the Comments

Table S3.5

Confusion Matrix for Partisan Prejudice (Comments) Classifier

	Predicted		
Actual	Non-Prejudiced	Prejudiced	Total
Non-Prejudiced	278	31	309
Prejudiced	25	66	91
Total	303	97	400

Note. This confusion matrix shows prediction accuracy of the partisan prejudice (comments) classifier. Rows represent actual class labels; columns represent predicted class labels. Diagonal values indicate correct predictions, while off-diagonal values show misclassifications.

S4: Interdependence of Partisan Prejudice and Genericity

Despite being theoretically distinct constructs, genericity and partisan prejudice are highly correlated in our training data. Most instances of generic expressions in the posts also convey negative affect toward a political outgroup. As a consequence, there is a risk that the partisan prejudice classifier learned to treat the presence of generic formulations as a strong proxy for polarization, not because genericity is definitionally equivalent to negativity, but because in our data these features co-occur so frequently. This could inflate the observed correlation between the two classifier outputs, conflating genuine theoretical overlap with artifacts of the training distribution. As described above, we tried to prevent this by augmenting training data with synthetic posts that vary one construct while holding the other one constant. To verify, that the partisan prejudice classifier does not capture variance that conceptually belongs to genericity, we constructed a contrastive test set. Specifically, we selected a sample of 100 posts (from the same population as the training data) and manually created parallel versions that differed only with regard to genericity, while keeping the level of partisan prejudice as well as the overall structure and content of the text as similar as possible. We then applied the classifiers to these parallel versions and compared the binary

predictions. If the partisan prejudice classifier is sensitive to genericity cues, we would expect systematically more prejudiced predictions for the generic versions.

Statistical comparisons were conducted using two-sided paired-samples *t* tests to assess whether classifier predictions differed systematically between the paired conditions. The genericity classifier successfully distinguished between the two versions, showing a very large effect size, $t(99) = 9.95$, $p < .001$, $d_z = 0.99$, 95% CI [0.81, 1.22]. In contrast, the partisan prejudice classifier did not significantly differentiate between the two versions, $t(99) = 1.00$, $p = .320$, $d_z = 0.10$, 95% CI [–0.10, 0.25]. While this result does not imply that the classifiers are entirely independent, it indicates that genericity—as manipulated in our contrastive test set—was captured much more effectively by the genericity classifier than by the partisan prejudice classifier.

S5: Correlations between Predictors

To inspect the bivariate relations between our predictors, we calculated zero-order correlations (see Table S5.1). Given the hierarchical structure of our data, we additionally conducted correlation analyses for the decomposed predictors. Specifically, we ran correlation analyses for the surface-level means of the predictors (see Table S5.2), as well as for the group-mean centered post-level predictors (see Table S5.3). Across both between- and within-surface correlations two correlations stood out: First, there was a relatively strong correlation between linguistic form and label target. The proportion of noun labels increased from 25% for individuals, to 63% for subgroups, to 97% for groups. Second, there was a strong correlation between genericity and partisan prejudice in the posts, with 5% of non-prejudiced posts being classified as generic versus 55% of prejudiced ones. These moderate correlations with our independent measures support our decision to include them as covariates, as they indicate potential confounding.

Zero-Order Correlations

Table S5.1

Zero-Order Correlations for Predictors

Variable	*M*	*SD*	1	2	3	4	5
1 Linguistic form (Noun)	0.59	0.49					
2 Genericity (Generic)	0.29	0.45	0.12***				
3 PP (Prejudiced)	0.35	0.48	0.31***	0.54***			
4 Label target (Subgroup)	0.20	0.40	0.04***	−0.19***	−0.04***		
5 Label target (Group)	0.37	0.48	0.59***	0.19***	0.34***	−0.38***	
6 Party (Republicans)	0.49	0.50	−0.18***	−0.10***	−0.11***	0.14***	−0.25***

Note. M = mean. *SD* = standard deviations. Values below the diagonal are phi coefficients for correlations between binary variables. Asterisks denote significance levels (*$p < .05$, **$p < .01$, ***$p < .001$).

Between-Surface Correlations

Table S5.2

Between-Surface Correlations for Predictors

Variable	*M*	*SD*	1	2	3	4	5
1 Linguistic form (Noun)	0.60	0.40					
2 Genericity (Generic)	0.37	0.41	0.28***				
3 PP (Prejudiced)	0.41	0.43	0.39***	0.60***			
4 Label target (Subgroup)	0.18	0.31	−0.01	−0.18***	−0.06***		
5 Label target (Group)	0.37	0.40	0.56***	0.36***	0.49***	−0.33***	
6 Party (Republicans)	0.45	0.41	−0.17***	−0.16***	−0.16***	0.10***	−0.24***

Note. M = mean. *SD* = standard deviations. Values below the diagonal are pearson correlations. Asterisks denote significance levels (*$p < .05$, **$p < .01$, ***$p < .001$).

Within-Surface Correlations

Table S5.3

Within-Surface Correlations for Predictors

Variable	*M*	*SD*	1	2	3	4	5
1 Linguistic form (Noun)	0.00	0.44					
2 Genericity (Generic)	0.00	0.35	0.06***				
3 PP (Prejudiced)	0.00	0.36	0.17***	0.38***			
4 Label target (Subgroup)	0.00	0.37	0.04***	−0.12***	−0.03***		
5 Label target (Group)	0.00	0.43	0.56***	0.12***	0.22***	−0.40***	
6 Party (Republicans)	0.00	0.44	−0.17***	−0.08***	−0.07***	0.11***	−0.22***

Note. M = mean. *SD* = standard deviations. Values below the diagonal are pearson correlations. Asterisks denote significance levels (*$p < .05$, **$p < .01$, ***$p < .001$).

S6: Angry Reactions

To investigate the effect of linguistic form and genericity on the number of angry reactions, we fitted multilevel negative binomial regressions with random intercepts for surfaces and fixed effects for linguistic form, genericity, affective polarization in the posts, label target (individuals vs. subgroup vs. group), and party, applying the same multilevel decomposition as for the confirmatory analyses. We again tested the effect of linguistic form using the full dataset (including posts about individuals and excluding genericity as a predictor) and the effect of genericity on a dataset excluding posts about individuals. Moreover, we included an offset for the logarithm of the post view count to account for differences in exposure, meaning that model coefficients reflect differences in the outcome per unit of exposure rather than absolute counts. To test for moderation effects, we again ran different models in which we incrementally added higher-order interaction terms and compared them using BIC—calculated following the recommendations by Lorah and Womack (2019). We report only the two-way interaction models as these had the best fit (lowest BIC) for both the linguistic form and genericity analyses. Importantly, model diagnostics using the DHARMa package (Hartig, 2024) revealed significant underdispersion (dispersion = 0.14, $p = .054$) and moderate overprediction of zeros (ratioObsSim = 0.59, $p < .001$), which may lead to slightly conservative hypothesis tests.

To analyze the effect of linguistic form on the rate of angry reactions, we fitted a series of models: Model 0 was an intercept-only model, Model 1 included only within- and between-surface linguistic form as a predictor, Model 2 additionally controlled for within- and between-surface label target (individual, subgroup, group), affective polarization, and party, and Model 3 included two-way interactions between within-surface predictors (see Table S6.1). The effect of within-surface linguistic form on the rate of angry reactions was not significant in Model 1, IRR = 1.01, 95% CI [0.97, 1.05], p = .574, but significant in Models 2, IRR = 1.06, 95% CI [1.02, 1.12], p = .009, and 3, IRR = 1.08, 95% CI [1.02, 1.13], p = .005. To illustrate, within a surface with the average proportion of noun labels (59%), a median post (~13,100 views) is expected to receive 11.4 vs. 12.2 angry reactions (or

0.87 vs. 0.94 angry reactions per 1,000 views), depending on whether the label is an adjective vs. a noun (based on the estimates from Model 3). However, the effect of linguistic form was moderated by label target, with stronger effects for both subgroups, IRR = 1.37, 95% CI [1.24, 1.52], $p < .001$, and groups, IRR = 1.30, 95% CI [1.15, 1.47], $p < .001$, compared to individuals. An additional pairwise comparison showed that the difference between subgroup and group was not significant, IRR = 0.95, 95% CI [0.81, 1.11], $p = .706$. Simple effects analyses (for a surface with the average proportion of 20% subgroup references and 37% group references) revealed significant positive effects of linguistic form for subgroups, IRR = 1.26, 95% CI [1.16, 1.37], $p < .001$, and groups, IRR = 1.19, 95% CI [1.07, 1.33], $p = .001$, but a significant negative effect for individuals, IRR = 0.92, 95% CI [0.86, 0.98], $p = .008$, indicating a reduced rate of angry reactions for noun compared to adjective labels.

To test whether genericity predicted the rate of angry reactions, we followed the same modelling approach. Across Models 1, 2, and 3, within-surface genericity significantly predicted the rate of angry reactions, indicating that posts using labels generically vs. non-generically had higher rates of angry reactions (see Table S6.2). The estimated effect for within-surface genericity was IRR = 1.20, 95% CI [1.12, 1.27], $p < .001$ in Model 1, IRR = 1.16, 95% CI [1.08, 1.24], $p < .001$ in Model 2, and IRR = 1.13, 95% CI [1.06, 1.22], $p < .001$ in Model 3. To illustrate, within a surface with the average proportion of generic posts (29%), a median post (~12,800 views) is expected to receive 14.2 vs. 16.2 angry reactions (or 1.12 vs. 1.27 angry reactions per 1,000 views), depending on whether the label is used non-generically vs. generically (based on the estimates from Model 3). The effect of genericity was not moderated by any of the other predictors.

Table S6.1

Random-Intercept Negative Binomial Models Predicting the Rate of Angry Reactions from Linguistic Form (N_{posts} *= 45,321;* $N_{surfaces}$ *= 3,846)*

	Model 0	Model 1	Model 2	Model 3
Surface-level variables				
% of noun labels		1.64 [1.38, 1.94]***	0.74 [0.59, 0.92]**	0.74 [0.59, 0.93]*
% of subgroup references			2.90 [2.25, 3.73]***	2.93 [2.28, 3.77]***
% of group references			1.80 [1.40, 2.32]***	1.80 [1.39, 2.31]***
% of prejudiced posts			2.28 [1.92, 2.72]***	2.30 [1.93, 2.74]***
% of Republican references			1.08 [0.91, 1.28]	1.08 [0.91, 1.28]
Post-level variables				
Linguistic form (Noun)		1.01 [0.97, 1.05]	1.06 [1.02, 1.12]**	1.08 [1.02, 1.13]**
Label target (Subgroup)			1.19 [1.14, 1.26]***	1.25 [1.19, 1.31]***
Label target (Group)			0.87 [0.82, 0.91]***	0.88 [0.82, 0.93]***
PP (Prejudiced)			1.36 [1.30, 1.42]***	1.36 [1.30, 1.43]***
Party (Republicans)			1.28 [1.23, 1.33]***	1.28 [1.23, 1.33]***
Linguistic form * Label target (Subgroup)				1.37 [1.24, 1.52]***
Linguistic form * Label target (Group)				1.30 [1.15, 1.47]***
Linguistic form * PP				0.92 [0.80, 1.05]
Linguistic form * Party				0.98 [0.88, 1.08]
Label target (Subgroup) * PP				1.12 [0.96, 1.29]
Label target (Subgroup) * Party				0.89 [0.80, 1.00]
Label target (Group) * PP				1.15 [1.00, 1.33]*
Label target (Group) * Party				1.26 [1.11, 1.42]***
PP * Party				0.94 [0.84, 1.05]
$ICC_{surfaces}$	.203	.198	.181	.181
BIC	369,312	369,300	368,705	368,716
$R^2_{conditional}$	.203	.200	.195	.195
$R^2_{marginal}$	.000	.002	.017	.018

Note. Incidence rate ratios (*IRR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

Table S6.2

Random-Intercept Negative Binomial Models Predicting the Rate of Angry Reactions from Genericity (N_{posts} *= 25,778;* $N_{surfaces}$ *= 2,765)*

	Model 0	Model 1	Model 2	Model 3
Surface-level variables				
% of generic posts		1.87 [1.55, 2.26]***	1.20 [0.93, 1.55]	1.20 [0.93, 1.54]
% of noun labels			1.11 [0.82, 1.51]	1.10 [0.81, 1.50]
% of group references			0.54 [0.43, 0.70]***	0.54 [0.42, 0.70]***
% of prejudiced posts			2.38 [1.88, 3.01]***	2.36 [1.86, 2.99]***
% of Republican references			1.16 [0.95, 1.41]	1.16 [0.96, 1.42]
Post-level variables				
Genericity (Generic)		1.20 [1.12, 1.27]***	1.16 [1.08, 1.24]***	1.13 [1.06, 1.22]***
Linguistic form (Noun)			1.20 [1.12, 1.29]***	1.13 [1.03, 1.25]*
Label target (Group)			0.70 [0.66, 0.74]***	0.70 [0.66, 0.74]***
PP (Prejudiced)			1.33 [1.25, 1.41]***	1.35 [1.26, 1.43]***
Party (Republicans)			1.36 [1.29, 1.44]***	1.36 [1.29, 1.44]***
Genericity * Linguistic form				0.83 [0.63, 1.09]
Genericity * Label target				1.15 [0.95, 1.39]
Genericity * PP				1.06 [0.91, 1.24]
Genericity * Party				0.99 [0.82, 1.19]
Linguistic form * Label target				0.88 [0.73, 1.06]
Linguistic form * PP				0.91 [0.73, 1.13]
Linguistic form * Party				0.93 [0.78, 1.10]
Label target * PP				1.03 [0.88, 1.21]
Label target * Party				1.44 [1.26, 1.64]***
PP * Party				0.96 [0.82, 1.11]
$ICC_{surfaces}$	.197	.186	.180	.181
BIC	214,780	214,722	214,329	214,387
$R^2_{conditional}$	.197	.190	.195	.197
$R^2_{marginal}$	.000	.005	.019	.020

Note. Incidence rate ratios (*IRR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

S7: Top Ten News Media Outlets

To assess differences between news media outlets, we calculated the percentages of noun labels and generic references for the ten largest news media outlets in our dataset (see Table S7). The analysis reveals substantial noun label usage across all outlets: even the outlet with the lowest percentage (HuffPost) still relied on noun labels in nearly half of its posts containing partisan references (41.5%). Use of generic expressions was generally less frequent. Most outlets used generic references in fewer than 10% of posts, with several approaching near-zero levels (e.g., New York Times: 1.6%, CBS News: 1.7%). However, Breitbart was a clear outlier, with 39.3% of posts using labels generically. This unusually high percentage may reflect a deliberate editorial choice, with generic language serving as a distinguishing stylistic feature.

Table S7

Linguistic Labeling by Top 10 News Media Outlets

News media outlet	*N*	Noun labels (in %)	$N_{subgroup/group}$	Generic posts (in %)
NEWSMAX	2426	48.1	1225	9.4
The Hill	2118	48.3	1066	3.7
Breitbart	1076	60.5	644	39.3
Fox News	1001	45.2	447	5.8
Washington Post	784	56.0	471	2.3
CBS News	588	48.6	296	1.7
ABC News	578	46.9	276	3.6
POLITICO	563	68.0	392	8.2
The New York Times	530	61.3	322	1.6
HuffPost	528	41.5	155	9.7

Note. The table shows the proportion of noun labels and generic posts for the top ten news media outlets. *N* = number of posts in our dataset. $N_{subgroup/group}$ = number of posts where label refers to a subgroup or group. The percentage of generic posts is based only on posts where the label refers to subgroups or groups. We included surfaces from the "Media/news company", "News & media website", "Broadcasting & media production company", and "Newspaper" categories. News media outlets were ranked by the number of posts contributed to the dataset.

S8: Moderation Analyses

To test for moderation effects, we repeated the confirmatory analyses but additionally included party (i.e., whether the label referred to Republicans or Democrats) and post-level interaction terms. To balance omitted variable bias and overfitting concerns, we ran different models in which we incrementally added higher-order interaction terms and compared them using BIC—calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. We report only the two-way interaction models as these had the best fit (lowest BIC) for both the linguistic form and genericity analyses.

We theorized that labeling effects are mediated by perceived entitativity: Labeling increases perceived entitativity, which in turn heightens partisan prejudice. However, perceived entitativity might also—or instead—*moderate* the effect of partisan prejudice expressed in the post. In this view, higher entitativity does not directly increase outgroup prejudice but amplifies the impact of partisan animosity in the post—for instance by strengthening individual-to-group generalization (Crawford et al., 2002). If labeling tends to co-occur with negative post content, such moderation could still produce the observed pattern of results. We examined this possibility using results from the moderation analyses. A moderating role would manifest in a significant interaction of labeling with partisan prejudice (with stronger labeling effects for prejudiced posts), a mediating role in a labeling main effect.

We again found significant positive effects of both linguistic form (OR = 1.18, 95% CI [1.14, 1.21], $p < .001$) and genericity (OR = 1.21, 95% CI [1.17, 1.25], $p < .001$) on partisan prejudice in the comments (see Tables S8.1 and S8.2). Importantly, neither effect interacted with partisan prejudice (linguistic form: OR = 0.94, 95% CI [0.88, 1.00], $p = .057$; genericity: OR = 0.97, 95% CI [0.91, 1.04], $p = .461$). Simple effects analyses revealed that both linguistic form and genericity significantly increased prejudice both for non-prejudiced posts (linguistic form: OR = 1.20, 95% CI [1.16, 1.25], $p < .001$; genericity: OR = 1.23, 95% CI [1.16, 1.29], $p < .001$) and for prejudiced posts (linguistic form: OR = 1.13, 95% CI [1.07, 1.19], $p < .001$; genericity: OR = 1.19, 95% CI [1.15, 1.25], $p < .001$). Contrary

to the moderation account, both effects were descriptively *weaker* for prejudiced posts. Overall, these findings are in line with the view that entitativity acts as a mediator rather than a moderator. This view is further supported by Graf et al.'s (2020) observation that labeling effects emerged regardless of message valence, as well as by studies that directly manipulated entitativity and documented effects on outgroup stereotypes and prejudice even in the absence of valenced information about the group (e.g., Castano, Sacchi, et al., 2003; Dang, Liu, Ren, & Gu, 2018).

Moreover, the moderation analyses showed the effect of linguistic form to be moderated only by label target (see Table S8.1), with stronger effects for both subgroups, OR = 1.09, 95% CI [1.03, 1.14], p = .002, and groups, OR = 1.16, 95% CI [1.09, 1.23], p < .001, compared to individuals. An additional pairwise comparison showed that the difference between subgroup and group was not significant, OR = 1.06, 95% CI [0.98, 1.15], p = .163. Simple effects analyses (for a surface with the average proportion of 20% subgroup references and 37% group references) revealed significant positive effects of linguistic form across all levels of label target (individuals: OR = 1.10, 95% CI [1.06, 1.13], p < .001; subgroup: OR = 1.19, 95% CI [1.14, 1.24], p < .001, group: OR = 1.27, 95% CI [1.20, 1.34], p < .001). The effect of genericity was not moderated by any other variable (see Table S8.2).

Table S8.1

Moderation Analysis for Linguistic Form Effect ($N_{comments}$ *= 769,684;* N_{posts} *= 45,780;* $N_{surfaces}$ *= 3,878;* $N_{commenters}$ *= 391,386)*

	No Interactions	Two-Way
Surface-level variables		
Proportion of noun labels	1.06 [0.98, 1.16]	1.07 [0.98, 1.16]
Proportion of subgroup references	1.77 [1.61, 1.94]***	1.77 [1.62, 1.94]***
Proportion of group references	1.71 [1.55, 1.88]***	1.70 [1.55, 1.87]***
Proportion of prejudiced posts	1.33 [1.25, 1.42]***	1.33 [1.25, 1.42]***
Proportion of Republican references	0.93 [0.87, 0.99]*	0.93 [0.87, 0.99]*
Post-level variables		
Linguistic form (Noun)	1.16 [1.13, 1.19]***	1.18 [1.14, 1.21]***
Label target (Subgroup)	1.54 [1.50, 1.58]***	1.55 [1.51, 1.59]***
Label target (Group)	1.34 [1.30, 1.37]***	1.32 [1.28, 1.36]***
PP (Prejudiced)	1.17 [1.15, 1.20]***	1.18 [1.15, 1.21]***
Party (Republicans)	1.06 [1.04, 1.08]***	1.06 [1.04, 1.08]***
Linguistic form * Label target (Subgroup)		1.09 [1.03, 1.14]**
Linguistic form * Label target (Group)		1.16 [1.09, 1.23]***
Linguistic form * PP		0.94 [0.88, 1.00]
Linguistic form * Party		1.01 [0.96, 1.07]
Label target (Subgroup) * PP		1.03 [0.96, 1.10]
Label target (Subgroup) * Party		1.06 [1.01, 1.13]*
Label target (Group) * PP		1.00 [0.93, 1.07]
Label target (Group) * Party		1.16 [1.09, 1.23]***
PP * Party		1.02 [0.96, 1.07]
$ICC_{surfaces}$	.037	.037
ICC_{posts}	.090	.090
BIC	775,579	775,604
$R^2_{conditional}$	.147	.146
$R^2_{marginal}$	.022	.023

Note. Odds ratios (*OR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces and posts. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

Table S8.2

Moderation Analysis for Genericity Effect ($N_{comments}$ *= 437,744;* N_{posts} *= 26,026;* $N_{surfaces}$ *= 2,785;* $N_{commenters}$ *= 248,334)*

	No Interactions	Two-Way
Surface-level variables		
Proportion of generic posts	0.88 [0.80, 0.96]**	0.88 [0.80, 0.96]**
Proportion of noun labels	1.05 [0.94, 1.17]	1.05 [0.94, 1.17]
Proportion of group references	0.97 [0.89, 1.07]	0.97 [0.89, 1.06]
Proportion of prejudiced posts	1.41 [1.29, 1.53]***	1.41 [1.29, 1.53]***
Proportion of Republican references	0.97 [0.91, 1.04]	0.98 [0.91, 1.04]
Post-level variables		
Genericity (Generic)	1.21 [1.17, 1.25]***	1.21 [1.17, 1.25]***
Linguistic form (Noun)	1.16 [1.12, 1.20]***	1.17 [1.11, 1.22]***
Label target (Group)	0.88 [0.85, 0.90]***	0.87 [0.85, 0.90]***
PP (Prejudiced)	1.08 [1.05, 1.12]***	1.08 [1.05, 1.12]***
Party (Republicans)	1.12 [1.10, 1.15]***	1.13 [1.10, 1.15]***
Genericity * Linguistic form		1.04 [0.92, 1.19]
Genericity * Label target		1.06 [0.97, 1.16]
Genericity * PP		0.97 [0.91, 1.04]
Genericity * Party		1.06 [0.98, 1.15]
Linguistic form * Label target		1.01 [0.92, 1.10]
Linguistic form * PP		0.98 [0.89, 1.09]
Linguistic form * Party		0.99 [0.91, 1.07]
Label target * PP		0.94 [0.87, 1.01]
Label target * Party		1.12 [1.05, 1.19]***
PP * Party		1.00 [0.93, 1.07]
$ICC_{surfaces}$	.032	.032
ICC_{posts}	.086	.086
BIC	480,023	480,099
$R^2_{conditional}$	.123	.123
$R^2_{marginal}$	.005	.005

Note. Odds ratios (*OR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces and posts. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

S9: Selection Effects

It is possible that our findings reflect selection effects rather than within-user change (see Mamakos & Finkel, 2023). That is, posts using generic noun labels may not increase users' partisan prejudice but may simply elicit comments disproportionately from users who are already highly prejudiced. In our main analyses, we did not include random intercepts for commenters because the large number of commenters (N = 391,386) and posts (N = 45,780) made crossed random effects computationally infeasible. To nonetheless assess potential selection effects, we conducted additional analyses on a filtered dataset in which sparse random-effect levels were excluded to reduce the number of factor-level combinations to a manageable size (< 50 million). This required removing commenters and posts with fewer than seven comments in the linguistic-form models and fewer than five comments in the genericity models.

When we repeated the analyses using the filtered dataset, both hypotheses remained supported with comparable effect sizes. For H1, within-surface linguistic form remained a significant positive predictor of partisan prejudice (Model 1: OR = 1.42, 95% CI [1.35, 1.49], p < .001; Model 2: OR = 1.13, 95% CI [1.05, 1.20], p < .001; see Table S9.1). For H2, within-surface genericity also remained a significant positive predictor (Model 1: OR = 1.18, 95% CI [1.11, 1.26], p < .001; Model 2: OR = 1.18, 95% CI [1.10, 1.27], p < .001; see Table S9.2). These findings are in line with the idea that linguistic labeling affects *how* users comment rather than *which* users comment.

Table S9.1

Results for Hypothesis 1 with Random Intercepts for Commenters ($N_{comments}$ *= 61,421;* N_{posts} *= 5,751;* $N_{surfaces}$ *= 39;* $N_{commenters}$ *= 2,829)*

	Model 0	Model 1	Model 2
Surface-level variables			
Proportion of noun labels		1.77 [0.82, 3.82]	8.36 [2.59, 27.01]***
Proportion of prejudiced posts			0.42 [0.25, 0.70]***
Proportion of subgroup references			0.80 [0.31, 2.07]
Proportion of group references			0.24 [0.08, 0.72]*
Post-level variables			
Linguistic form (Noun)		1.42 [1.35, 1.49]***	1.13 [1.05, 1.20]***
PP (Prejudiced)			1.09 [1.01, 1.18]*
Label target (Subgroup)			1.70 [1.59, 1.81]***
Label target (Group)			1.42 [1.31, 1.53]***
$ICC_{surfaces}$	.011	.008	.000
ICC_{posts}	.080	.075	.069
$ICC_{commenters}$	.136	.138	.141
BIC	68,463	68,290	68,066
$R^2_{conditional}$	.227	.226	.224
$R^2_{marginal}$	.000	.007	.018

Note. Odds ratios (*OR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces, posts, and commenters. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

Table S9.2

Results for Hypothesis 2 with Random Intercepts for Commenters ($N_{comments}$ *= 70,733;* N_{posts} *= 8,068;* $N_{surfaces}$ *= 204;* $N_{commenters}$ *= 5,482)*

	Model 0	Model 1	Model 2
Surface-level variables			
Proportion of generic posts		0.93 [0.74, 1.17]	0.74 [0.47, 1.17]
Proportion of noun labels			0.99 [0.48, 2.06]
Proportion of prejudiced posts			1.30 [0.86, 1.96]
Proportion of group references			0.77 [0.54, 1.10]
Post-level variables			
Genericity (Generic)		1.18 [1.11, 1.26]***	1.18 [1.10, 1.27]***
Linguistic form (Noun)			1.12 [1.05, 1.19]***
PP (Prejudiced)			1.07 [1.01, 1.14]*
Label target (Group)			0.82 [0.78, 0.86]***
$ICC_{surfaces}$	.010	.010	.009
ICC_{posts}	.054	.054	.053
$ICC_{commenters}$	.154	.154	.155
BIC	83,352	83,342	83,316
$R^2_{conditional}$	.219	.219	.219
$R^2_{marginal}$	.000	.001	.003

Note. Odds ratios (*OR*) with 95% confidence intervals in brackets. Asterisks denote significance levels (*p < .05, **p < .01, ***p < .001). *ICC* = intra-level correlation coefficient for surfaces and posts. *BIC* = Bayesian information criterion calculated following Lorah and Womack's (2019) recommendations for multilevel models, using level-specific penalties. $R^2_{conditional}$ = variance explained by both fixed and random effects and $R^2_{marginal}$ = variance explained by fixed effects only (Nakagawa & Schielzeth, 2013).

S10: Ingroup Versus Outgroup Labels

We argued that partisan prejudice results from outgroup labeling and entitativity. However, heightened ingroup entitativity can strengthen ingroup identification (Castano, Yzerbyt, et al., 2003; Hogg et al., 2007; Sacchi et al., 2009; Yzerbyt et al., 2000) and may therefore also explain heightened partisan prejudice (Brandt & Vallabha, 2025; Iyengar et al., 2019; Mason, 2015). Although we cannot directly rule out ingroup effects, they seem less likely given that partisan prejudice in comments should overwhelmingly target (a) outgroups rather than ingroups, and (b) the group referenced in the post. An additional analysis examining comments containing explicit partisan references ("Republicans", "Democrats", or "Dems") supports the latter assumption. Of all prejudiced comments 31.1% explicitly referred to a single party (66.6% referred to no party and 2.3% referred to both parties). Of single-party references, 71.4% referred to the party mentioned in the post. Since users should be more likely to use explicit references when targeting a party not mentioned in the post, the correspondence for comments with implicit references is likely much higher. Although these considerations do not rule out ingroup effects, they suggest such effects are less likely captured in our design.

References

Brandt, M. J., & Vallabha, S. (2025). Intraindividual changes in political identity strength (but not direction) are associated with political animosity in the United States and the Netherlands. *Personality and Social Psychology Bulletin*, *51*(5), 828–844. https://doi.org/10.1177/01461672231203471

Castano, E., Sacchi, S., & Gries, P. H. (2003). The perception of the other in international relations: Evidence for the polarizing effect of entitativity. *Political Psychology*, *24*(3), 449–468. https://doi.org/10.1111/0162-895X.00336

Castano, E., Yzerbyt, V., & Bourguignon, D. (2003). We are one and I like it: The impact of ingroup entitativity on ingroup identification. *European Journal of Social Psychology*, *33*(6), 735–754. https://doi.org/10.1002/ejsp.175

Crawford, M. T., Sherman, S. J., & Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology*, *83*(5), 1076–1094. https://doi.org/10.1037/0022-3514.83.5.1076

Dang, J., Liu, L., Ren, D., & Gu, Z. (2018). "Groupy" allies are more beneficial while "groupy" enemies are more harmful. *Social Psychological and Personality Science*, *9*(8), 925–934. https://doi.org/10.1177/1948550617729409

Graf, S., Linhartova, P., & Sczesny, S. (2020). The effects of news report valence and linguistic labels on prejudice against social minorities. *Media Psychology*, *23*(2), 215–243. https://doi.org/10.1080/15213269.2019.1584571

Hartig, F. (2024). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models* (Version 0.4.7) [Computer software]. https://CRAN.R-project.org/package=DHARMa

Hogg, M. A., Sherman, D. K., Dierselhuis, J., Maitner, A. T., & Moffitt, G. (2007). Uncertainty, entitativity, and group identification. *Journal of Experimental Social Psychology*, *43*(1), 135–142. https://doi.org/10.1016/j.jesp.2005.12.008

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Mamakos, M., & Finkel, E. J. (2023). The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus*, *2*(10), Article pgad325. https://doi.org/10.1093/pnasnexus/pgad325

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Sacchi, S., Castano, E., & Brauer, M. (2009). Perceiving one's nation: Entitativity, agency and security in the international arena. *International Journal of Psychology*, *44*(5), 321–332. https://doi.org/10.1080/00207590802236233

Yzerbyt, V., Castano, E., Leyens, J.-P., & Paladino, M.-P. (2000). The primacy of the ingroup: The interplay of entitativity and identification. *European Review of Social Psychology*, *11*(1), 257–295. https://doi.org/10.1080/14792772043000059

APPENDIX
Declarations

Author Contributions Statement

The author contributions for all three dissertation articles are reported using the Contributor Roles Taxonomy (CRediT) and listed below in Table 1. All authors have reviewed and agreed to this author contributions statement.

Table 1

Author Contributions Based on CRediT Roles

	First Article		Second Article		Third Article	
Role	**CS**	**JD**	**CS**	**JD**	**CS**	**JD**
Conceptualization	×	×	×	(×)	×	
Data curation	×		×		×	
Formal analysis	×		×		×	
Funding acquisition						
Investigation	×		×		×	
Methodology	×	×	×	(×)	×	
Project Administration	×		×		×	
Resources						
Software					×	
Supervision		×		×		×
Validation	×	×	×	×	×	
Visualization	×		×		×	
Writing – original draft	×		×		×	
Writing – review & editing	×	×	×	×	×	×

Note. The table shows author contributions based on the Contributor Roles Taxonomy (CRediT). Roles that did not apply to a given project were left blank. CS = Carsten W. Sander, JD = Juliane Degner.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Erklärung gemäß *(bitte Zutreffendes ankreuzen)*

☐ **§ 4 (1c) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 5 (4d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich,

Carsten Sander (Vorname, Nachname),

dass ich mich an einer anderen Universität oder Fakultät noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer Doktorprüfung bemüht habe.

Hamburg, 13.04.2026
Ort, Datum

C. Sander
Unterschrift

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR PSYCHOLOGIE UND
BEWEGUNGSWISSENSCHAFT
Institut für Bewegungswissenschaft
Institut für Psychologie

Eidesstattliche Erklärung nach *(bitte Zutreffendes ankreuzen)*

☐ **§ 7 (4) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☒ **§ 9 (1c und 1d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich an Eides statt,

1. dass die von mir vorgelegte Dissertation nicht Gegenstand eines anderen Prüfungsverfahrens gewesen oder in einem solchen Verfahren als ungenügend beurteilt worden ist.
2. dass ich die von mir vorgelegte Dissertation selbst verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die wörtlich oder inhaltlich übernommenen Stellen habe ich als solche kenntlich gemacht.

Hamburg, 13.04.2026
Ort, Datum

C. Sander
Unterschrift