

Benefit or Backlash? Examining the Consequences of Expanding Definitions of Prejudice

August 18, 2026

Abstract

Definitions of prejudice have expanded in recent political discourse, yet the downstream consequences remain unclear. This paper develops a theory of backlash to expanding definitions, arguing that when individuals encounter allegations they feel are exaggerated or false, they become more skeptical of subsequent allegations and less concerned about the relevant form of prejudice. We test this in the context of antisemitism in the U.S., using an original survey experiment ($N = 5,110$) exposing respondents to an allegation labeling a pro-Palestinian protest as antisemitic. Exposure reduces belief in the validity of a subsequent allegation and lowers assessments of antisemitism's severity in society, supporting a backlash interpretation. These effects appear driven largely by a motivated reasoning process aimed at protecting moral self-image. While our empirical focus is antisemitism, the theory and results are broadly significant for scholars of racism and practitioners engaged in definitional debates over what constitutes prejudice.

Word Count: 11,696

1 Introduction

Accusations of prejudice play an important role in our democratic politics. They help citizens identify social harms, establish moral boundaries, and can help to catalyze institutional responses. From civil rights litigation to more recent campus debates, labeling a behavior as racist or prejudiced signals that a norm has been violated and that corrective action may be warranted. In this sense, accusations of prejudice function as a useful tool by shaping how citizens interpret events and assess the severity of social problems.

However, in recent years, the concepts used to describe bias and prejudice (e.g., racism, sexism, antisemitism, etc.) have broadened significantly, extending from their original definitions of overt and intentional hostility toward often more subtle or ambiguous acts and attitudes (Haslam 2016). This conceptual expansion reflects genuine moral progress in many ways, but it may also carry potential negative consequences. As definitions of prejudice become more expansive, so does public disagreement about their application. Are off-color jokes racist? Is certain criticism of Israel antisemitic? Is holding the door for a woman sexist? These questions lack self-evident answers, and the allegations that answer them in the affirmative have become a divisive issue of contemporary American political discourse.

Although this conceptual inflation is widely recognized, its consequences remain poorly understood. The optimistic expectation, drawing from a robust literature in political science, suggests that persuasive messages reliably shift beliefs, even on contested topics (Bishin et al. 2016; Coppock 2023). The implication is that exposure to expanding concepts of prejudice should move people toward accepting the validity of new accusations and heighten their concern about the underlying form of prejudice.

The alternative possibility is backlash. This paper develops and tests a theory of backlash to accusations that stem from expanded definitions of prejudice. Drawing on Aesop's fable of "the Boy Who Cried Wolf," we argue that when individuals encounter allegations of prejudice that they find unpersuasive or exaggerated, those encounters do not merely fail to persuade, they actively undermine the credibility of subsequent allegations and dampen overall concern about the prejudice

as a societal problem. Much like the fabled shepherds who stopped heeding the boy's warnings, a non-trivial portion of the public may come to discount genuine or more serious claims of prejudice after repeated exposure to allegations they find unconvincing. Because people will be more likely to encounter allegations of prejudice they distrust when the definitions of prejudice expand, we expect conceptual inflation to frequently trigger a backlash effect.

We test this theory in the context of antisemitism in the U.S., using an original survey experiment of over 5,000 U.S. adults. Antisemitism is an especially well-suited test case because it combines a highly salient definitional debate, a substantial share of the public that regards certain allegations as overreaching, and high-profile organizations that regularly issue accusations of prejudice in response to political events like pro-Palestinian protests. The interest of this project is not to adjudicate the validity of particular claims of antisemitism; rather, we are concerned with understanding their political and psychological consequences on the American public. As such, the theory that we develop in this paper is not specific to antisemitism. We expect that analogous dynamics have the potential to arise anytime allegations of prejudice become highly contested, and we view this paper as establishing the general framework and providing the first empirical evidence of this backlash effect.

Our results show that exposure to a contentious allegation of antisemitism decreases belief in later, unrelated allegations and reduces perceptions of antisemitism as a serious societal problem, especially among respondents who support pro-Palestinian protests. Exploratory mechanism tests indicate that the backlash effect is most likely to occur among individuals who are morally implicated by the contentious allegation—that is, those who feel targeted by the allegation. Backlash appears more common when skepticism is motivated by a desire to protect oneself from a perceived moral threat than when it is driven by unmotivated disbelief in the allegation, though we cannot rule out the possibility that unmotivated disbelief is also a driver.

These findings underscore that the consequences of broadening definitions of prejudice are not merely semantic, but also political and psychological. Accusations that extend beyond widely shared definitions risk provoking backlash, which can erode the credibility of future claims and

diminish public concern for genuine harms—not by increasing prejudice directly, but by shaping how individuals interpret subsequent acts of prejudice. This backlash effect therefore results in later incidents being judged as less severe or less credible than they otherwise would have been.

The remainder of the paper proceeds as follows. We first outline our theoretical framework, and then explain how it operates in the context of antisemitism. Next, we describe our experimental design and measurement strategy, and present our results. Finally, we test mechanisms, adjudicate alternative explanations, and discuss broader implications for understanding how expanding definitions of prejudice can generate political and psychological backlash.

2 Theory

2.1 Prejudice and Concept Creep

Over the past several decades, definitions of prejudice have expanded significantly. While prejudice in the mid-1900s was largely conceptualized as intergroup antipathy (Allport 1954, Dixon et al. 2012), today we argue over whether accidentally misgendering someone is transphobic, whether compliments can be racist, and whether remedies to discrimination constitute forms of prejudice. This expansion of definitions of prejudice is not a mere popular perception: psychologists have systematically documented the phenomenon under the label of “concept creep” (Haslam 2016)—the expansion of harm-related concepts to cover milder and more ambiguous instances (“vertical creep”), or to new contexts or classes of phenomena entirely (“horizontal creep”) (Dakin et al. 2023).

An example of concept creep in the realm of prejudice concerns the widening definition of what constitutes racism. In the latter half of the 20th century, scholars departed from narrower conceptualizations to include more subtle ideologies such as “racial resentment,” “modern racism,” or “symbolic racism” (McConahay and Hough Jr. 1976; Kinder and Sears 1981; Kinder and Sanders 1996) as well as unconscious forms of negative affect such as implicit bias and “aversive racism” (Dovidio and Gaertner 1986). In the early 2000s, some scholars attempted to further

expand the definition of racism to include microaggressions, defined as “subtle insults (verbal, non-verbal, and/or visual) directed towards people of color, often automatically or unconsciously” (Pierce 1974; Solorzano et al. 2000; Sue et al. 2007). And in recent years, a maximally expansive, albeit controversial, definition of racism has emerged, wherein anyone who is not actively anti-racist is classified as racist (Kendi 2019).

Beyond racism, concept creep has occurred in other forms of prejudice as well.¹ “New antisemitism” for example, comprises not only hostility towards Jews, but also certain criticism of Israel or of Zionism (Foxman 2003; Judaken 2008; Waxman et al. 2022). “Benevolent sexism” expands the traditional definition of sexism to include certain positive, yet potentially patronizing, attitudes towards women (Glick and Fiske 1996), and the contemporary connotation of “homophobia” can refer to much more than the clinical fear it was coined to describe (Weinberg 1972; Haslam 2016). In short, concept creep has substantially broadened definitions of prejudice. Such linguistic expansions are not universally accepted (Herek 2000; Tetlock 1994; Concerned Jewish Faculty and Staff 2024), yet they have created conditions under which allegations of prejudice often refer to less severe or less obvious forms of prejudice than those to which these labels were traditionally confined.

2.2 Concept Creep and Backlash

The consequences of concept creep have been widely theorized, though rarely conclusively demonstrated. One possibility is that semantic shifts constitute a “meaningful expansion,” allowing previously unrecognized or ignored forms of prejudice to be brought to light, enabling accountability and understanding (Cascardi and Brown 2016). Indeed, a recent body of political science literature on persuasive messaging suggests that most of the time, messages move individuals’ opinions in the intended direction (Bishin et al. 2016; Coppock 2023). And in the context of prejudice, previous work has argued that voters may turn against politicians making implicit racial

¹Beyond the realm of prejudice, political science has been concerned with the “conceptual stretching” of terms like “democracy” for decades (Sartori 1970; Collier and Mahon 1993; Collier and Levitsky 1996; Gerring 1999).

appeals once prominent actors publicly expose their racial content (Mendelberg 2001).

But concept creep scholars have also suggested that expanding definitions of prejudice might induce backlash rather than acceptance (Haslam 2016; Haslam et al. 2020). For instance, Haslam et al. (2020) argue that “the result of this stretching may be an increasing lack of consensus on moral issues and a widening penumbra of cases that are morally ambiguous or contestable. The fact that people tend to understand harm so differently may therefore help to explain some yet to be determined proportion of public disagreement on morally charged social issues.” Related literature on implicit bias and prejudice documents such initial rejection empirically (Howell and Ratliff 2017; Vitriol and Moskowitz 2021; Lofaro et al. 2024). For example, when individuals are told after an implicit bias test that they harbor implicit racial bias, they often derogate the test or dismiss the accusation rather than accept it (Howell and Ratliff 2017; Vitriol and Moskowitz 2021).

Importantly, we hypothesize that this backlash is *not* simply limited to the contested cases themselves. More worryingly, expanding the definition of prejudice might induce backlash against and distrust in the prejudice concept itself, particularly when ambiguous cases come to be seen as emblematic and typical. Dakin et al. (2023) refer to this as “semantic dilution” and warn it could “lead people to perceive harm-related concepts as less urgent.” Scholars of racism have worried about such “conceptual inflation” for decades (Miles 1989; Blum 2002), and adjacent research on framing suggests that civil rights frames may lose their resonance when deployed too widely (Snow and Corrigall-Brown 2005; Yazdiha 2017; Carlson 2025).

This underlying logic—that distrust generated by one allegation can spread to color judgments of unrelated future ones—is reminiscent of Aesop’s fable of “The Boy Who Cried Wolf.” Like the shepherds in the story, some modern citizens may be lulled into a false sense of security through exposure to claims they consider to be misleading. When expanded definitions of prejudice result in people encountering claims of prejudice that they believe to be false (either through motivated or unmotivated reasoning), they may come to distrust similar claims in the future and become less concerned about that prejudice as a social problem.

Such a dynamic closely resembles a Bayesian process of belief revision. In a Bayesian frame-

work, individuals hold prior beliefs about the frequency or seriousness of prejudice and update those beliefs in light of new information (Gerber and Green 1999; Bullock 2009). If they encounter an accusation of prejudice that they judge to be weak, exaggerated, or inappropriate, they may rationally revise their posterior beliefs downward, concluding that such prejudice is less prevalent or severe than they had previously believed.

From this perspective, an implausible accusation might lead people to recalibrate their expectations about how often prejudice is present in interpersonal contact and communication, or revise their beliefs of its societal significance. If the initial accusation is seen as representative of a broader pattern of overstatement, Bayesian updating should also lead to more generalized skepticism about subsequent claims (Gerber and Green 1999). This generalized skepticism should appear most strongly in reaction to *vague* subsequent claims, as individuals judging the veracity of new allegations must rely more heavily on their priors when details about the incident are scarce than when they are plentiful.

Importantly, this process of downward belief revision should be concentrated among individuals who distrust the initial allegation (who we call “skeptics”). People who trust claims of prejudice made under expanded definitions (“believers”) have no reason to begin distrusting future allegations and should update, if at all, in the opposite direction after encountering an initial allegation they judge to be true.

Dakin et al. (2023) provides the first empirical evidence of backlash to concept creep, although not in the context of prejudice. They demonstrate that survey respondents exposed to “marginal” examples of bullying and workplace harassment became less concerned about these problems, and expressed less willingness to donate to charities combating these issues. Crucially, this backlash was not limited to other “marginal” cases; respondents became less concerned about bullying and harassment as societal problems *in general*. To our knowledge, no existing literature tests this hypothesis in the context of prejudice, but we argue that theoretical similarities of these cases lend support to the extension of the backlash hypothesis to prejudice.

3 Context: Antisemitism in the United States

We test our theory in the context of antisemitism in the United States, an area marked by a long-standing and highly salient definitional debate (Foxman 2003; Klug 2003; IHRA 2016; Waxman et al. 2022; Nexus Project 2024; Kaleem 2025; Stempel 2025). In a textbook example of concept creep, the Anti-Defamation League (ADL)—the U.S.’s most prominent organization dedicated to combating antisemitism—formally expanded its definition of antisemitism in 2024 to include certain acts and speech critical of Zionism or of Israel (Anti-Defamation League 2024a).² This definitional expansion allowed the ADL to regularly issue allegations of antisemitism in response to events such as pro-Palestinian protests³ (Anti-Defamation League 2024b), despite widespread public disagreement over whether pro-Palestinian slogans or criticisms of Israel should be classified as antisemitism.⁴ This combination of frequent allegations made by elites and substantial public disagreement over their legitimacy makes antisemitism a particularly well-suited context to test our theory.

In this paper, we do not take a position on whether criticism of Israel or anti-Zionist activism *truly* constitutes antisemitism.⁵ Instead, we view the current public debate as a unique opportunity

²As of 2024, the ADL’s new approach to Israel-related expressions “comports with the IHRA definition of antisemitism” (Anti-Defamation League 2024a). Some examples that the IHRA definition would consider antisemitic include: (1) “Denying the Jewish people their right to self-determination, e.g., by claiming that the existence of a State of Israel is a racist endeavor” and (2) “Drawing comparisons of contemporary Israeli policy to that of the Nazis” (IHRA 2016).

³For example, a glance at the ADL’s H.E.A.T. Map reveals numerous incidents labeled by the ADL as “Antisemitic Incident:Harassment” which occurred at “anti-Israel rallies” (Anti-Defamation League 2024b). The incidents include protesters chanting or holding signs reading “From the river to the sea Palestine will be free” (Incident ID 71816), “Resistance is not terrorism” (Incident ID 73001), “Zionism = racism” (Incident ID 72506), and similar slogans.

⁴For example, Wright et al. (2024) surveyed non-Jewish students on U.S. college campuses and found that 34% believe that the statement “Israel does not have the right to exist” is not antisemitic while 76% believe that the statement “Israel violates human rights of the Palestinian people” is not antisemitic. Similarly, Pape (2024) found that 74% of U.S. college students did not interpret “From the river to the sea, Palestine will be free” as a call for violence. Specifically, 26% believed that the slogan was a call for expulsion or genocide against Israeli Jews, while 43% believed in a non-violent interpretation and 31% did not know what the phrase meant.

⁵For readers who are interested in reading more about this expansive debate, we especially recommend the Waxman et al. (2022) piece *Arguing about antisemitism*. For those interested in the contours of contemporary antisemitism more generally, we recommend Hersch and Royden (2022). More broadly, readers may wish to consult the following: Beller (2015); Dinnerstein (1995); Frindte et al. (2005); Antoniou et al. (2020); Bassi (2023); Berkman (2021); Bilewicz (2022); Cohen et al. (2009); King and Weiner (2007); Kofta et al. (2020); Kaplan and Small (2006); Meuleman et al. (2019); Rosenfeld (2019); Solomos (2017); Stanislawski (2017). These sources are not intended to constitute an exhaustive list of works, but rather serve as a representative cross section of the extensive literature on this subject.

to test how people respond to expanded definitions of prejudice. Accordingly, we are interested not in adjudicating the validity of specific allegations, but rather in understanding how people respond to allegations of prejudice that are not universally accepted. Our focus is on how credible these claims are perceived to be, and on the psychological and political effects that exposure to them may produce. We theorize that skeptics of contentious allegations will respond by discounting future allegations and down-weighting their perceived severity of antisemitism in the U.S. today, resulting in a backlash effect.

In sum, antisemitism provides a near-ideal test case for our theory because it combines a) a salient example of concept creep, b) high-profile organizations issuing contested accusations, and c) a substantial number of skeptics. We therefore propose the following two pre-registered hypotheses:

Hypothesis 1 (H1) *Skeptics exposed to a contentious allegation of antisemitism will be less likely to judge subsequent claims of antisemitism as genuinely antisemitic.*⁶

Hypothesis 2 (H2) *Skeptics exposed to a contentious allegation of antisemitism will express less concern about antisemitism as a societal problem.*

4 Methods

To test our hypotheses, we embedded an experiment in a nationally representative survey of Americans ($N = 5,110$). We pre-register our experimental design with Center for Open Science. We explicitly note any deviations as they arise and provide further details in Supplementary Information Section A.11. In our experiment, all respondents are assigned to either a) a treatment condition where they are informed that a Pro-Palestine rally has been accused of antisemitism, or b) a control condition where respondents are informed about the rally, but do not view the allegation of antisemitism. We then measure a range of outcomes for all respondents.

⁶As outlined above and in our pre-analysis plan, subsequent *vague* allegations constitute our main test for Hypothesis 1, with specific subsequent allegations representing a possible extension.

We now provide an abbreviated description of the experimental procedure, and include a full description of the exact survey flow and question wordings in Supplemental Information Section A.2. After providing demographic information, respondents were randomly assigned to exposure to a contentious claim of antisemitism. *All respondents* were informed about a real “Pro-Palestine rally,” at which protesters carried signs reading “Stand with Palestinian people’s resistance” and “Stop the lies against Palestinian resistance.” Afterwards, respondents *in the treatment group* viewed a social media post denouncing the protests as antisemitic and describing antisemitism as the protests’ “defining feature.” The post was presented as being made by a fictional nonprofit organization that characterizes itself as combating antisemitism (Figure 1). This treatment design attempts to mirror contentious claims of antisemitism as many may encounter them in the real world: as a passionate response to recent events on social media.⁷ Immediately after viewing the post, we ask *respondents in the treatment group* to assess the antisemitic nature of the rally in question on a five-point scale from “Definitely not antisemitic” to “Definitely antisemitic”.⁸

Following the introduction to the protest and the treatment group’s exposure to (and assessment of) the contentious allegation, *all respondents* reported their perceptions of antisemitism as a societal problem. First, they were asked to rate on a scale of from 0 to 10 how much they think antisemitism is a problem in the United States today, where 0 indicates not a problem at all and 10 indicates antisemitism is the biggest problem. Respondents also rank-ordered a number of potentially important problems facing the country, with antisemitism being one of the featured problems.⁹ The scale question aims to capture the magnitude of respondents’ view of antisemitism as a problem, while the rank order question situates antisemitism in relation to other potential important problems that people may be concerned about. We use each of these outcomes to test Hypothesis 2. While tapping the same underlying concept, these measures differ in stringency:

⁷We derive the treatment from a combination of statements made by Jonathan Greenblatt, CEO and national director of the ADL, on his personal social media account. Specifically, we draw from statements posted on 7 November 2023, 22 April 2024, and 9 June 2024 on the social media platform X. Additionally, the incidents described in the treatment and the additional allegation outcomes correspond to real incidents listed as antisemitic by the ADL.

⁸While it would be ideal to ask all respondents this question, it is not feasible to do so for the control group without treating them.

⁹The other potential important problems respondents placed into rank order alongside antisemitism were Islamophobia, inflation, crime, poverty, climate change, and childcare costs.

Figure 1: Contentious Claim of Antisemitism Displayed to Respondents in Treatment Group A non-profit organization focused on combating hate and antisemitism denounced these protests as antisemitic. The following tweet details their response:



Nonprofit Combatting Antisemitism

@nonprofitcombattingantisemitism



Make no mistake, these "protests" against Israel are a sign of resurgent antisemitism, shameless in its transparency. These protests weren't just marred by antisemitism -- it was their defining feature.

12:00 PM · Apr 22, 2024

the 0–10 scale provides a direct assessment of perceived severity, while the rank-order question measures test whether treatment effects extend to comparative judgments.

To test the effect of accusations stemming from expanded definitions on respondents' belief in or agreement with subsequent claims of antisemitism, we solicit reactions to three additional incidents from all respondents. After a brief description of each incident, the survey informed every respondent that “a non-profit organization focused on combating hate and antisemitism denounced the incident as antisemitic.” These additional allegations outcomes were designed such that they would be less contentious than the allegation described in the treatment. The first, which we refer to as the “Left-wing” allegation, reports an incident where the words “Free Gaza” were spray painted onto a synagogue. The second, “Right-wing” allegation, describes a situation where individuals posted fliers reading “six Jewish companies own 96% of the world’s media.” The last incident, the “Overt” allegation, describes someone drawing a swastika and the words “Don’t be Jewish” on a mailbox. Following each incident’s description, respondents indicated their perceptions of the incidents’ antisemitic natures on the same 5-point scale used in the treatment skepticism question

(“Definitely not antisemitic” to “Definitely antisemitic”).

A fourth outcome did not recount the particulars of any event, but rather referred to a vague incident denounced by a nonprofit organization. For respondents in the control [treatment] group, the prompt read, “You’ve heard about a lot of different circumstances that have been described as antisemitic: [the Pro-Palestine rally,] the vandalism at the synagogue, the flier about media companies, and the swastika on the mailbox. Without knowing any specific details about a fourth [fifth] circumstance that was denounced as antisemitic, would you guess that the circumstance was...” and asked respondents to grade the vaguely-described incident on the same 5-point scale from “definitely not” to “definitely” antisemitic. Because even control respondents had viewed several less contentious allegations of antisemitism prior to this question, this outcome provides a sort of placebo test, isolating the effect of the contentious allegation from any general response to allegation exposure.¹⁰ Observing a treatment effect on respondents’ perceptions of the antisemitic nature of this vague incident is a direct test of Hypothesis 1. If respondents discount the validity of a future allegation because they were previously exposed to an allegation they did not agree with, this would constitute a backlash effect.

We also probe possible heterogeneous results based on the reason for rejecting the treatment allegation. While a purely Bayesian framework anticipates backlash from unmotivated skeptics, it is also possible that backlash effects are stronger when motivated reasoning accounts for the rejection of the initial accusation. This account predicts stronger treatment effects among certain groups of respondents. Following feedback from trusted colleagues, we made modifications to some of our pre-specified mechanism tests. As such, modified mechanism tests should be treated as exploratory. We adjudicate between mechanisms by examining which groups drive the treatment effects.

¹⁰Because we ask about ratings and rankings of antisemitism as a problem before the additional uncontentious allegations of antisemitism, we do not have a comparable placebo for these outcomes. Power constraints prevent us from using a true three-arm design. We further discuss this point in Supplemental Information Section A.2.

4.1 Study Population and Estimation Strategy

We administered our survey from July 16 to July 24, 2025 in partnership with Cint, a survey firm which draws respondents from a network of verified online panels. Recruitment is designed to be a nationally representative survey of Americans.¹¹ As noted above and specified in our pre-registered hypotheses, we only expect a backlash effect among skeptics (i.e., those who distrust the treatment allegation). No such backlash is expected among those who trust the initial allegation (believers). As such, our quantity of interest is the subgroup average treatment effect among skeptics.

Since the treatment allegation is only shown to the treatment group, we cannot measure skepticism directly for all respondents. We address this problem by providing three complementary estimation strategies. First, we present the nationally representative, full-sample average treatment effect (ATE), which should theoretically be a conservative estimate of the quantity of interest as it includes potentially null or countervailing effects among believers. Second, we report our pre-registered test: the subgroup ATE among respondents whose answers to a validated pretreatment subset question indicate that they are likely to be skeptics.¹² Third, we report an analysis using principal stratification (Frangakis and Rubin 2002; Ding and Lu 2017; Hartman and Huang 2024), which more directly isolates the treatment effect among respondents who would be skeptical of the allegation if treated. We provide full details on the method of principal stratification and its identifying assumptions in Supplemental Information A.9. Our principal stratification analyses are not pre-registered.

These approaches all have their advantages and disadvantages: the full-sample analysis avoids subgroup classification entirely but consequently cannot isolate the theoretically relevant subgroup; the proxy-based analysis follows our pre-registered strategy but measures skepticism imperfectly; and the principal stratification analysis more directly targets the latent subgroup estimand but requires additional identifying assumptions. Consistent results across the three approaches would therefore provide convergent and robust evidence for the hypothesized backlash effect.

¹¹The sample composition deviates from our pre-analysis plan. We explain in Supplemental Information A.11.2 that the composition differs because of an error on the part of the survey firm.

¹²See Supplemental Information A.8 for further details on this strategy.

Finally, we estimate all treatment effects using linear regressions. As discussed in Supplemental Information A.10, the results do not meaningfully differ if we instead use ordered probit regressions for outcomes measured on continuous or quasi-continuous scales.

5 Results

5.1 A Backlash Effect

We begin by reporting our main result for Hypothesis 1. When asked to imagine a future incident labeled as antisemitic by a nonprofit organization that combats hate, treated respondents were less likely to judge the future incident as antisemitic. As shown in Table 1, we find this effect not only among the subset of likely skeptics (where it was theorized to occur), but also in our full sample of respondents. This effect also occurs in spite of respondents being only recently exposed to the three additional less-contentious claims of antisemitism from purported hate-combating organizations and despite the fact that respondents have undoubtedly encountered many contentious claims in their everyday lives over the last several months.¹³ One short and subtle encounter with a contentious allegation is enough to push respondents to discount subsequent accusations. Therefore, our study finds strong support for Hypothesis 1.

Figure 2 visualizes the magnitude of the negative treatment effect on perceptions of the vague allegation by translating the full-sample ATE coefficient (the smallest of the three coefficients shown in Table 1) into differences in predicted probabilities. We obtain these predicted probabilities from an ordered probit version of the full-sample model. Each point plots the difference in likelihood of a treated respondent viewing the Vague Future allegation at each level of the 5-point ordered scale relative to a respondent in the control group. The bars extending out from each

¹³It is plausible that our experimental intervention is best interpreted as an additional dose of a treatment that most respondents have already received. If the marginal effect of each additional dose declines with cumulative exposure, then our estimates provide a lower bound on the effect of a first (baseline) dose, but unbiased estimates of the marginal effect of increasing dosage around the current average level of exposure in the population. The cumulative effect of all doses individuals have received over time may therefore be substantially larger than the effects we report here. Future research should directly investigate this possibility.

Table 1: Treatment Effects on Perceptions of Vague Future Allegation

	Full Sample	Skeptics (Pre-Registered) Only	Skeptics (Principal Stratification) Only
Treatment	-0.074** (0.027)	-0.143*** (0.033)	-0.280*** (0.031)
Constant	3.098*** (0.092)	3.143*** (0.106)	3.065*** (0.097)
Controls	Yes	Yes	Yes
N (effective)	5,058	3,543	3,808
R^2	0.068	0.036	0.052

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variable is measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an antisemitic incident. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.23 in the Supplementary Information presents the coefficients for all control variables.

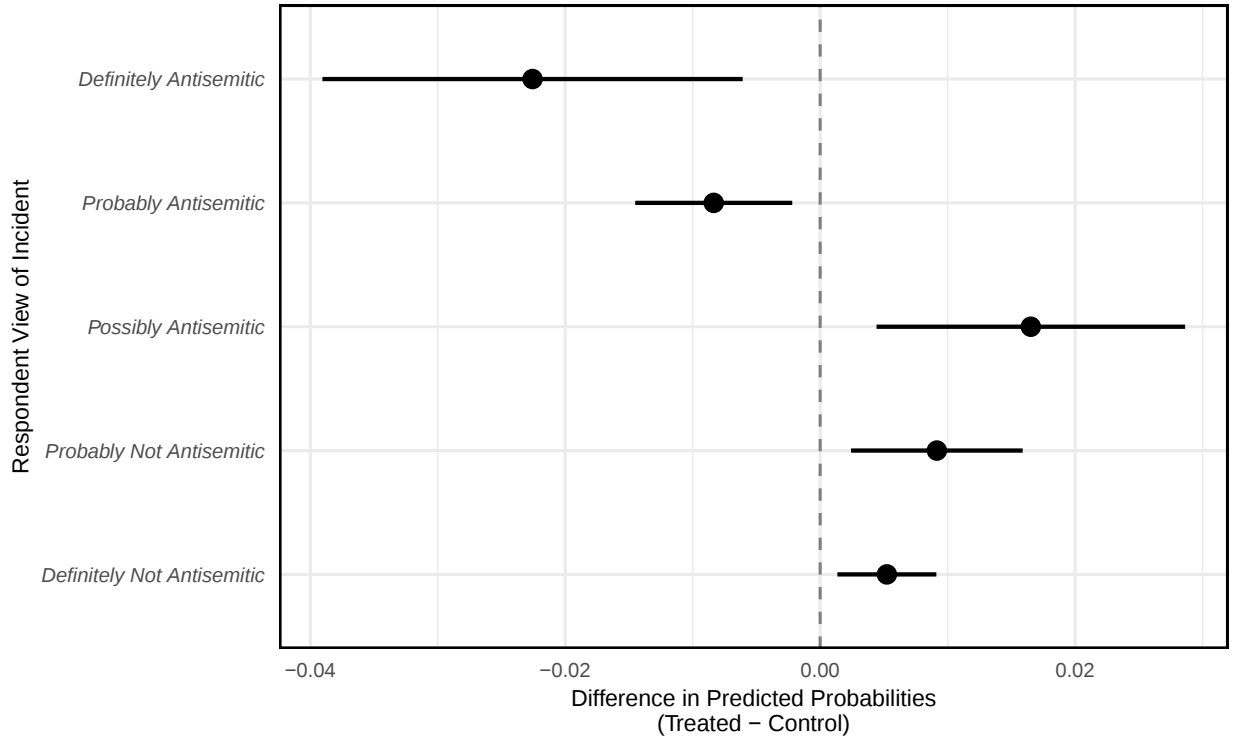
point represent 95% confidence intervals on the differences in probabilities. Treated respondents were more likely to perceive the future incident as definitely not, probably not, and possibly antisemitic. Correspondingly, they were less likely to indicate the incident was probably and definitely antisemitic. All differences in probability are statistically significant, suggesting that treatment meaningfully shifts the likelihood of respondents choosing each category on the ordered scale.

We believe this result—which appears consistently across all three estimation strategies—provides evidence of the hypothesized backlash effect.¹⁴ An encounter with a distrusted accusation of antisemitism leads skeptics to become less likely to believe that a subsequent accusation describes an antisemitic incident.

We next evaluate how respondents judge more specific antisemitic incidents after treatment exposure, displaying the results in Table 2. Here, the results from the pre-registered and principal stratification estimation strategies diverge. The results of the pre-registered specification indicate that treated skeptics judge these incidents to be only slightly less antisemitic than skeptics in the control group, while the principal stratification results suggest that the differences are both

¹⁴As discussed in more detail in Supplemental Information Section A.7, a non-negative effect among skeptics would be highly unlikely even when only considering the negative full-sample ATE. The consistent results across all three estimation strategies make this possibility vanishingly small.

Figure 2: Predicted Probabilities of Different Responses to Vague Allegation Across Treatment Groups



statistically and substantively significant. The null full-sample ATE is consistent with either finding: a near-zero estimate here could be a function of null effects among both skeptics and believers or of countervailing effects which cancel each other out.

Regardless, both interpretations suggest that treated skeptics become directionally less likely to believe subsequent, specific and unrelated, claims of antisemitism following exposure to an allegation they do not believe, and that any effect on these outcomes is smaller in magnitude than the effect we observe for the “Vague Future Allegation” outcome. This combination of results is consistent with our proposed Bayesian updating mechanism, as respondents do not need to rely as heavily on (new) priors to evaluate specific post-treatment allegations as they do to evaluate vague allegations.¹⁵

¹⁵We do note, however, that we find null results on a separate pre-registered test of Bayesian updating mechanism. We develop a novel measure of propensity to engage in Bayesian updating: we ask respondents whether screens cause eye damage, then inform them of recent studies which find they do not, and then ask the question again. We call those who update in the correct direction Bayesians, and those who do not non-Bayesians. We interact this variable with treatment, and find null effects in Supplemental Information Section A.3. We view the most likely explanation of this

Table 2: Treatment Effects on Perceptions of Other Allegations of Antisemitism

	Full Sample	Skeptics (Pre-Registered) Only	Skeptics (Principal Stratification) Only
<i>“Left-wing” Allegation</i>			
Treatment	-0.006 (0.031)	-0.054 (0.038)	-0.240*** (0.035)
Constant	2.829*** (0.105)	2.865*** (0.120)	2.794*** (0.111)
N (effective)	5,068	3,551	3,814
R^2	0.119	0.051	0.058
<i>“Right-wing” Allegation</i>			
Treatment	-0.019 (0.033)	-0.066 (0.039)	-0.204*** (0.036)
Constant	2.860*** (0.110)	2.862*** (0.124)	2.819*** (0.115)
N (effective)	5,064	3,545	3,810
R^2	0.080	0.054	0.058
<i>“Overt” Allegation</i>			
Treatment	-0.036 (0.030)	-0.043 (0.037)	-0.162*** (0.035)
Constant	3.178*** (0.100)	3.170*** (0.118)	3.187*** (0.109)
N (effective)	5,055	3,538	3,804
R^2	0.109	0.106	0.114
Controls	Yes	Yes	Yes

Note:

Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variables are measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an actually antisemitic incident. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Tables A.24, A.25, and A.26 in the Supplementary Information present the coefficients for all control variables.

5.2 Exposure to Treatment Leads to Lower Perceptions of Antisemitism as a Societal Problem

A downstream effect of treatment exposure is that it pushes skeptics to rate antisemitism as less of a problem in the United States today than they would absent treatment exposure. Treatment exposure does not, however, unambiguously translate into viewing antisemitism as less of a problem

null result to be a failure in the question. Our eye damage question is not a validated measure of propensity to engage in Bayesian updating, and this sort of overt, immediate updating task might differ quite substantially from the more subtle background updating implied by our theory.

compared to other issues. Table 3 reports the effects of treatment on the two outcomes that ascertain respondents' perceptions of the severity of antisemitism using all three estimation strategies.

Table 3: Treatment Effects on Perceptions of Antisemitism as a Problem

	Full Sample	Skeptics (Pre-Registered) Only	Skeptics (Principal Stratification) Only
<i>Severity of Problem (0–10 Scale)</i>			
Treatment	-0.152* (0.070)	-0.192* (0.084)	-0.497*** (0.079)
Constant	4.637*** (0.235)	4.771*** (0.271)	4.505*** (0.251)
N (effective)	5,072	3,553	3,818
R ²	0.076	0.069	0.067
<i>Problem Rank Out of 7 Issues</i>			
Treatment	0.007 (0.046)	-0.026 (0.054)	-0.196*** (0.051)
Constant	2.689*** (0.155)	2.908*** (0.173)	2.711*** (0.159)
N (effective)	5,044	3,532	3,795
R ²	0.067	0.023	0.021
Controls	Yes	Yes	Yes

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. Severity of problem is measured on a 0 to 10 scale, with higher values indicating a respondent views antisemitism as a more serious problem. Problem rank is a respondent's ranking of antisemitism against six other issues, reverse-coded so that it runs from 1 to 7, with higher values indicating a respondent ranks antisemitism as more important relative to those issues. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Tables A.27 and A.28 in the Supplementary Information present the coefficients for all control variables.

Across all three strategies, we detect a statistically significant effect of treatment on skeptics' rating of antisemitism as a societal problem on a 0-10 scale. Substantively, the effect suggests that skeptics exposed to treatment rated the severity of the problem between 0.15 and 0.50 points lower on the 0-10 scale. Treatment exposure as minimal as a single social media post is enough to trigger a detectable backlash in which skeptics (and respondents as a whole) view antisemitism as less of a problem.

At the same time, exposure to treatment did not yield a consistent detectable effect on the rank respondents gave antisemitism in comparison to other potential societal problems. Similar to Table

2, the results are significant in the principal stratification model and null in the pre-registered model. Again, the null full-sample ATE is compatible with either being correct. Taken together, these findings suggest that treatment exposure can modestly reduce concern about antisemitism but may leave broader perceptions of its importance relative to other problems unchanged. It is, however, important to note that the rank order outcome is a substantially harder test, especially in a treatment saturated environment such as ours. It requires substantial thought and more careful consideration from respondents to rank problems in relation to one another as opposed to rating one problem on a 0-10 scale. Therefore, one possible reason for null effects is that when people are asked to consider the severity of antisemitism more deeply, encountering a treatment allegation has less of an effect. By contrast, the 0-10 scale is more intuitive and so may more directly reflect respondents' initial instincts.

Importantly, the significant results in the full sample for both the 0–10 rating and the “Vague Future Allegation” outcomes indicate that the *net* effect of the treatment allegation is a decrease in both perceived severity of antisemitism and belief in other allegations among the general public. There are two possible explanations for this result: a) the backlash effect among skeptics is larger in magnitude than any countervailing effect among believers, or b) the two effects are equal in magnitude, but the larger share of skeptics in the population creates a negative full-sample ATE. While we cannot definitively rule out either explanation, we show in Supplemental Information A.9 that our results are more consistent with the latter.

5.3 Replication in Supplemental Survey

As additional evidence, we fielded a supplemental survey among a sample of undergraduates from five U.S. universities (N = 756) in the Fall of 2025.¹⁶ Despite the vastly smaller sample size

¹⁶The supplemental survey was substantively identical to the main survey save for three deviations. First, an alternative specification of the treatment was used, in which the control group received no mention of the pro-Palestinian protests. Second, the supplemental survey omitted all pre-treatment variables due to priming concerns. Third, the supplemental survey was embedded at the end of a larger omnibus survey fielded among U.S. undergraduates. We use demographic and ideological variables that happened to be included in the earlier portions of the omnibus survey as controls in our analysis. Because we lack the necessary pre-treatment variables to replicate the pre-registered analysis, we present only full-sample results in Table 4. See Supplemental Information Section A.4 for additional

and differing sample composition, we observe results that are substantively similar and larger in magnitude than in the main survey. Table 4 displays full-sample ATEs for the main outcomes, and additional details of the survey are included in Section A.4 of the Supplemental Information. The larger magnitudes in the supplemental survey are consistent with our theoretical framework, as undergraduates tend to be more sympathetic to the pro-Palestine movement than the general public and thus are more likely to be skeptical of the treatment allegation (Pape 2024). These results suggest that under optimal conditions (a more susceptible population and an arguably stronger treatment¹⁷), backlash can be stronger than the effect we observe in the general population survey.

Table 4: Treatment Effects on Main Outcomes from a Supplemental Undergraduate Survey

	Undergraduate Omnibus Sample	
	Vague Future Allegation	Severity of Problem (0-10 Scale)
Treatment	-0.659*** (0.056)	-0.332* (0.162)
Constant	3.784*** (0.163)	6.864*** (0.469)
Controls	Yes	Yes
Observations	730	726
R^2	0.256	0.195

Note: Treatment effects are derived from OLS models that also control for respondent gender, religion, race/ethnicity, partisan identification, self-reported ideology, and whether the respondent views Israel negatively. The first dependent variable is measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an antisemitic incident. The second dependent variable is measured on a 0 to 10 scale, with higher values indicating a respondent believes antisemitism is a greater societal problem. Respondent age and education are not included as controls because every respondent in this sample is an undergraduate. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.29 in the Supplementary Information presents the coefficients for all control variables.

details.

¹⁷See Supplemental Information A.2 for details on why this treatment might be stronger than the one used in the main survey.

5.4 Other Outcomes

In pre-registered exploratory analyses, discussed in further detail in Supplemental Information A.6, we test the effects of treatment exposure on several additional outcomes, including endorsement of antisemitic and Islamophobic tropes, attitudes towards free speech, and willingness to donate to a charity focused on combating antisemitism. Two findings of interest emerge from these analyses. First, across all estimation strategies, we find no evidence that skeptics themselves become more antisemitic in response to treatment. The donation outcome, however, sees mixed results. As shown in Table 5, the pre-registered models finds no significant difference in donation willingness between treated and control skeptics, while the principal stratification model suggests that treatment reduces donation amounts by about \$2.56 among skeptics and increases donation amounts by about \$4.76 among believers. Together, these results suggest that treatment exposure may (but does not definitively) reduce willingness to donate among skeptics. However, any effect that may exist among skeptics is offset by a countervailing effect among believers.

Table 5: Effect of Treatment on Willingness to Donate to a Charity Combatting Antisemitism

	Full Sample	Skeptics Only		Believers Only	
		Pre-Registered	Principal Stratification	Pre-Registered	Principal Stratification
Treatment	0.171 (1.013)	-0.579 (1.179)	-2.561* (1.121)	1.894 (1.959)	4.762*** (1.440)
Constant	33.757*** (3.451)	33.560*** (3.799)	32.757*** (3.584)	42.237*** (8.428)	39.123*** (4.635)
Controls	Yes	Yes	Yes	Yes	Yes
N (effective)	4,093	2,872	3,060	1,214	2,163
R ²	0.064	0.065	0.034	0.085	0.100

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variable is the amount, in US dollars, that a respondent pledged to give to the Foundation to Combat Antisemitism if they were to win an in-survey raffle. It ranges from \$0 to \$100, with higher values indicating a respondent willing to give more. In order to be shown this question, respondents first had to click “Yes” to a question asking if they would like to be entered into a raffle, which is why the number of observations is lower for this outcome than for other outcomes. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.31 in the Supplementary Information presents the coefficients for all control variables.

6 Does the Reason for Initial Skepticism Matter?

Our argument thus far presumes that any skepticism of the initial accusation is sufficient to trigger an updating process that leads to backlash. But it is possible that *motivation* for skepticism matters. Drawing on the motivated reasoning literature, we distinguish between two classes: “accuracy” motivations and “directional” motivations (Kunda 1990). Whereas accuracy motivations drive people to seek the truth, directional motivations compel individuals to process information in a way that arrives at a desired conclusion. Existing literature demonstrates that such motivated reasoning behavior is widespread (Lord et al. 1979; Taber and Lodge 2006; Gaines et al. 2007; Kahan 2013; Bolsen and Palm 2019). Previous literature has considered a wide variety of directional motivations: avoiding unpleasant tasks (Balcetis and Dunning 2010), avoiding believing oneself to be ill (Ditto and Lopez 1992), and maintaining consistency with one’s social network (Bayes and Druckman 2021).

Although accuracy motivations alone might be sufficient to lead many people to be skeptical of the treatment allegation, we argue that at least two directional motivations might be relevant as well. First, individuals possess a strong directional motivation to maintain a positive moral self-image (Ditto et al. 2009). Individuals are strongly motivated to view themselves as moral, fair, and egalitarian actors (Steele 1988; Aquino and Reed 2002; Jordan et al. 2011), and challenges to this self-concept can trigger self-protective responses. When accusations of prejudice, even implicitly, imply that one’s views or allies are complicit in bias, individuals can experience moral dissonance: a tension between their self-perception as being good and fair and information implying moral complicity (Monin and Jordan 2009). To alleviate this discomfort, they may engage in a motivated reasoning process aimed at restoring their sense of moral adequacy rather than seeking factual accuracy. In our context, individuals who feel *targeted* by accusations of antisemitism may reject those accusations to preserve their moral self-image.

Second, individuals might experience a directional motivation to defend their social group. Social identity theory, developed by Tajfel and Turner (1979), offers an important lens for understanding how group membership and identification influence individuals’ perceptions and behaviors.

This theory holds that individuals derive part of their sense of self based on their group memberships. Individuals are thus motivated to maintain a positive image of the groups to which they belong in order to maintain a positive self image. While our experimental subjects are not members of the Palestinian in-group in the strictest sense, they may strongly identify with the pro-Palestinian movement, particularly when they perceive it as a moral cause (Heere and James 2007). As a result, when confronted with accusations of antisemitism that are linked to this in-group, they may experience a psychological need to protect and defend their in-group identity. One might think of this as a special case of moral self-image defense, where the reason respondents' moral self-image is threatened is because their in-group-image is threatened.

Bayesian updating might be particularly likely when rejection of the initial accusation is directionally motivated.¹⁸ Research from neuroscience suggests that motivated judgments are accompanied by distinctive affective processes in the brain, suggesting that they may differ from unmotivated judgments in character as well as origin (Westen et al. 2006). We thus view directional motivations as constituting a distinct causal pathway: one possibility is that treatment induces unmotivated skepticism, which in turn induces skepticism about future incidents; another is that treatment induces motivated skepticism, which could also induce skepticism about future incidents. Indeed, we argue that the effect of motivated skepticism on future skepticism might be larger than the effect of unmotivated skepticism.¹⁹ Scholars have found that beliefs that stem from motivated processes are affectively charged (Lodge and Taber 2005) and more resistant to change (Taber and Lodge 2006). Motivated skepticism might be internalized more deeply and become more salient for respondents, giving such skepticism a stronger pull on their priors. The mediated pathway running through motivated skepticism might therefore lead to stronger updating and account for a larger effect than

¹⁸Some might object to this account, arguing that we mix a rational theory of behavior (Bayesian updating) with a cognitive theory of behavior (motivated reasoning). We believe there is no inconsistency here. Our description of a Bayesian updating process is not intended to be literal; respondents do not calculate precise probabilities of antisemitism, but they do in some sense allow new information to alter their other beliefs. There is no particular reason to believe that such updating should only occur in response to information acquired in an unmotivated manner.

¹⁹In Supplemental Information Section A.1, we fully elucidate our causal framework using directed acyclic graphs. We view unmotivated skepticism and motivated skepticism as competing mediators of the effect of treatment on future skepticism. The full causal model includes other secondary-stage mediators of the effect of motivated skepticism on future skepticism. We do not discuss those here because they are substantively similar and observationally equivalent in our data to the simple mediated pathway we propose.

the mediated pathway running through unmotivated skepticism.

The following subsections present analyses testing these mechanisms. While our pre-analysis plan proposes a set of mechanism tests, we no longer view the pre-registered tests as the most informative approach and instead present a different set of exploratory tests. We clarify however that our main hypotheses and interpretation of the main effects did not differ between pre-registration and analysis. The mechanism tests we present in the following sections should be seen as purely exploratory, although certain tests still match our pre-analysis plan. We report results from our pre-registered mechanism tests in Supplementary Information Section A.11.3.²⁰

6.1 Moral Self-Defense

The clearest test of whether or not morally-motivated skepticism leads to stronger backlash than unmotivated skepticism would be to isolate which respondents rejected the initial accusation for motivated reasons. Unfortunately, we cannot directly observe the cognitive processes that lead to initial rejection. Instead, we rely a moderation approach. We argue that respondents' support for pro-Palestine protests should moderate the morally motivated skepticism pathway. For protest supporters, accusations that pro-Palestine protests are antisemitic plausibly implicate the supporters of the protests themselves in antisemitism. For non-protest supporters, accusations have little moral relevance. We therefore expect that protest support should moderate the relationship between treatment and backlash, *if morally motivated skepticism plays a causal role*.

We test this moderation effect by interacting treatment with support for pro-Palestine protests (measured pretreatment). Because the purpose of our mechanism tests is to discover which respondents drive treatment effects in general, we begin by reporting results from the full sample. In Table 6, we confirm that protest support significantly increases treatment effects. Backlash is stronger among those who support the protests. Not only is protest support associated with stronger

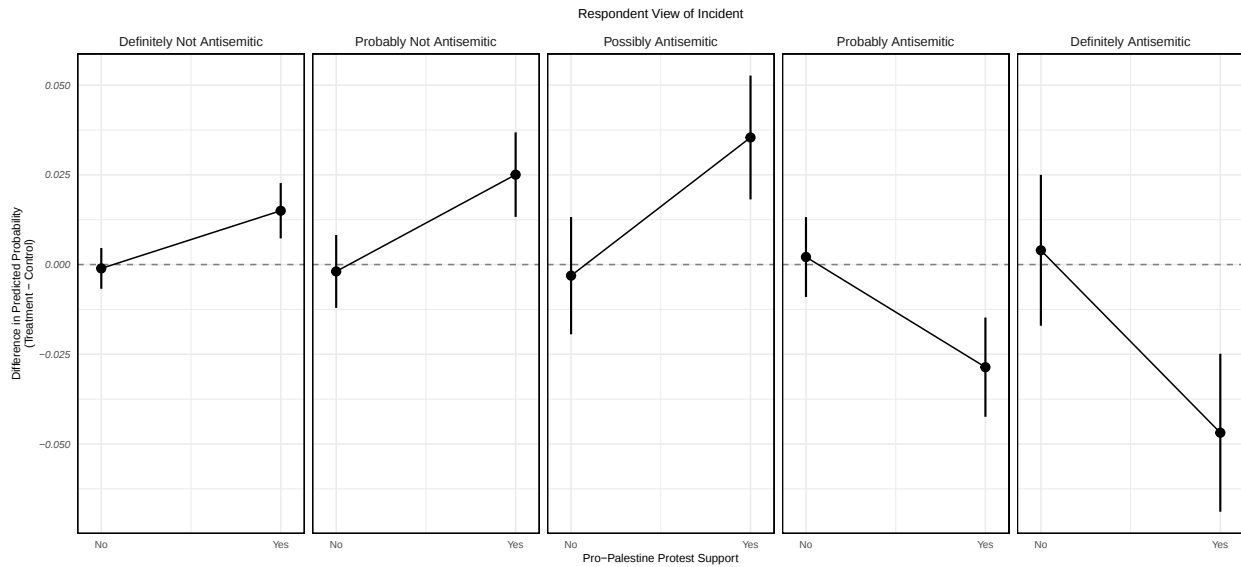
²⁰Conclusions from the two sets of mechanism tests differ slightly; following our pre-analysis plan for mechanism testing would lead us to conclude that moral self-defense plays a key causal role, *independently* of Bayesian updating. Our current analysis finds a role for both. Interested readers can find a full explanation in Supplementary Information Section A.11.3.

Table 6: Mechanism Test: Pro-Palestine Protest Support

	Vague Future Allegation	Severity of Problem
	Full Sample	Severity of Problem (0-10 Scale)
Treatment	0.012 (0.037)	-0.009 (0.095)
Supports Pro-Palestine Protests	0.042 (0.041)	0.250* (0.104)
Treatment $\times$ Protest Support	-0.187*** (0.055)	-0.303* (0.140)
Constant	3.086*** (0.095)	4.537*** (0.244)
Observations	5,039	5,049
R^2	0.071	0.077

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variable in the first column is measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an actually antisemitic incident. The dependent variable in the second column ranges from 0 to 10, with higher values indicating that the respondent believes antisemitism is more of a societal problem. Pro-Palestine protest support is measured as a binary, with a value of 1 indicating a respondent's support for the protests. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.49 in the Supplementary Information presents the coefficients for all control variables.

Figure 3: Differences in Predicted Probabilities of Rating the Vague Allegation by Protest Support



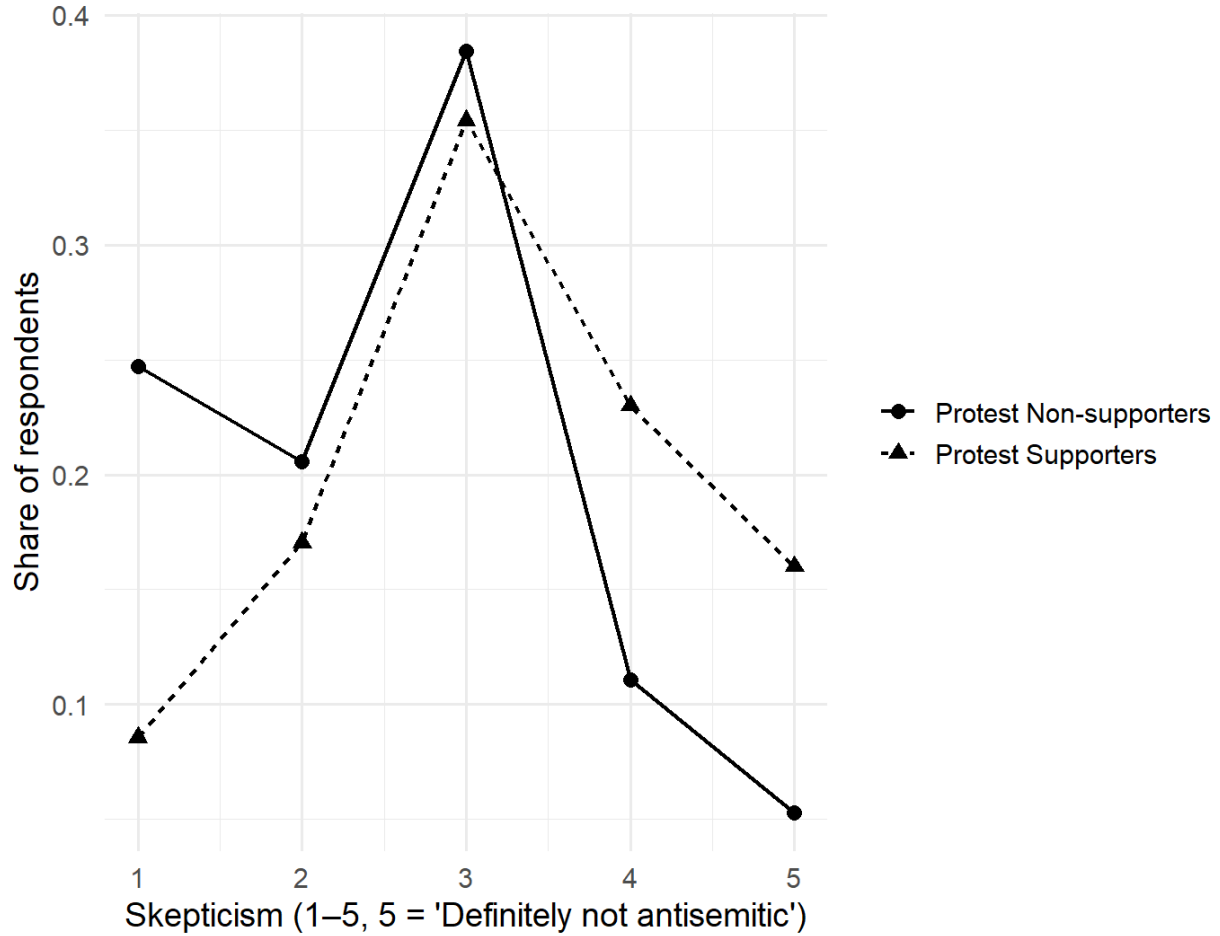
backlash, but treatment may not induce backlash among non-protest supporters at all. Figure 3 clearly demonstrates that we observe no treatment effects at all among non-protest supporters. As with Figure 2, we obtain these predicted probabilities from full-sample probit models.

Caution is warranted, however. Because we present this interaction in the full sample, it is possible that protest support simply proxies for skepticism. Most problematically, protest support may moderate not only the relationship between treatment and motivated skepticism, but also the relationship between treatment and *unmotivated* skepticism. Protest supporters may simply be more trusting of participants in the protest, leading them to view it as less likely that such protests are antisemitic for entirely unmotivated reasons. Even if morally motivated skepticism played no additional role, it is therefore possible that we would still observe a significant interaction effect between treatment and protest support in the full sample.

To test this alternative interpretation of the interaction, we plot skepticism of the treatment allegation among protest supporters and non-supporters. For this descriptive analysis, we subset to only treated respondents because we cannot observe treatment skepticism among respondents in control. We plot the results in Figure 4, where a 1 indicates a respondent called the treatment allegation “definitely antisemitic” and a 5 indicates a respondent viewed the allegation as “definitely not antisemitic.” We see that supporters were significantly more likely to think the treatment allegation was not antisemitic. This pattern suggests that protest support may induce skepticism, and we cannot say if this skepticism is motivated or unmotivated.

The pattern in Figure 4 suggests that the interaction between treatment and protest support is not necessarily sufficient to show that morally motivated skepticism plays an important causal role. To strengthen that argument, we test for treatment effects among skeptics who do not support the protests. We expect that protest support is a near perfect moderator of the effect of treatment on motivated skepticism; respondents who do not even support the protests should not feel targeted by the claim that these protests are antisemitic. But protest support is unlikely to be a *perfect* moderator of unmotivated skepticism — we should be able to identify some respondents who reject the allegation of antisemitism despite having no personal connection to the protests.

Figure 4: Share of Supporters vs. Non-Supporters at Each Level of Treatment Skepticism



To do so, we make use of our pre-registered skeptics-only specification (respondents who sympathize with Palestine). We do not use the principal stratification specification here because additional subsets of principal strata may rely on additional identifying assumptions and are rarely if ever employed in the principal stratification literature. We thus subset the group of pre-registered skeptics into a smaller subgroup based on their answer to the protest support question, isolating only those who do *not* support the protests. As this subgroup ($N = 1,519$) consists of likely skeptics who do not support the protests, we expect respondents in this subgroup to be largely *unmotivated* skeptics, who do not feel targeted by the treatment accusation but nonetheless do not fully believe it.²¹

²¹It is possible that although our pre-registered proxy variable generally does a good job of separating skeptics from believers, it may be less effective at classifying *non-supporter* skeptics. To validate whether or not this group is really skeptical of the treatment, we examine how the treated members of this subgroup react to the treatment allegation.

Table 7: Treatment Effects Among Non-Protest Supporting Likely Skeptics

	Non-Protest Supporting Likely Skeptics	
	Vague Future Allegation	Severity of Problem (0-10 Scale)
Treatment	-0.050 (0.050)	0.002 (0.128)
Constant	2.871*** (0.160)	4.210*** (0.409)
Observations	1,514	1,519
R^2	0.065	0.074

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variable in the first column is measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an actually antisemitic incident. The dependent variable in the two column ranges from 0 to 10, with higher values indicating that the respondent believes antisemitism is more of a societal problem. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.50 in the Supplementary Information presents the coefficients for all control variables.

As shown in Table 7, find no significant treatment effect among this group, strengthening the moral account. To alleviate concerns that a null result here may simply reflect low power, we show in Supplemental Information Section A.3 that subsetting to an even smaller group of respondents who *identify as members of* the pro-Palestine movement still yields highly significant results on the vague future allegation outcome. We still cannot entirely rule out the possibility that these non-supporter sympathizers are “weak skeptics,” and that a stronger degree of unmotivated skepticism is necessary to trigger backlash. But we view these null results as most consistent with the notion

We find that 74.73% are indeed skeptical, meaning they deem the treatment allegation to be one of the following: “definitely not antisemitic”, “probably not antisemitic”, or “possibly antisemitic”. Some readers might object to this classification, arguing that those who refer to the treatment allegation as “possibly antisemitic” should not be considered skeptics. In a separate pilot survey, we ask respondents in binary terms to classify the treatment allegation as antisemitic or not. We find that 66% of respondents say they do not believe the protests were antisemitic. In our main survey, only 27% of treated respondents say the protests were definitely or probably not antisemitic, while an additional 37% say the allegation was possibly antisemitic (for a total of 64%). We therefore feel reasonably confident that most respondents who answer “possibly antisemitic” would have given a skeptical response to a binarized question, indicating that they are indeed skeptics who should revise their priors downward. Further, principal stratification results confirm significant negative treatment effects in the “possibly antisemitic” stratum (see Supplemental Information A.9.4), suggesting that they behave as our theory predicts skeptics should.

that morally motivated skepticism constitutes a (if not the) core driver of the backlash we observe.

We supplement this evidence with an additional test of whether treatment effects are larger for respondents with a stronger motivation to maintain a positive moral self-image. We measure moral identity centrality using the moral identity internalization scale, a validated measure of the importance of morality to a person's maintenance of a positive self-image (Mulder and Aquino 2013; Aquino and Reed 2002). Like protest support, we expect that moral identity centrality should moderate the causal pathway running through morally motivated skepticism but not the one running through unmotivated skepticism.

Table 8: Mechanism Test: Moral Identity Centrality

	Vague Future Allegation		Severity of Problem (0-10 Scale)	
	Full Sample	Palestine Sympathizers Only	Full Sample	Palestine Sympathizers Only
Treatment	-0.071** (0.027)	-0.143*** (0.033)	-0.147* (0.070)	-0.194* (0.084)
Moral Identity Centrality (Standardized)	0.157*** (0.020)	0.151*** (0.024)	0.159** (0.050)	0.136* (0.060)
Treatment × Moral Identity	-0.062* (0.027)	-0.067* (0.033)	-0.052 (0.070)	-0.067 (0.084)
Constant	3.218*** (0.093)	3.249*** (0.107)	4.761*** (0.238)	4.861*** (0.273)
Observations	5,053	3,541	5,067	3,551
R ²	0.083	0.050	0.079	0.071

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variable in the first two columns is measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an actually antisemitic incident. The dependent variable in the second two columns ranges from 0 to 10, with higher values indicating that the respondent believes antisemitism is more of a societal problem. Moral identity centrality is standardized, such that the coefficients reflect the predicted effect of a one standard deviation increase. The original moral identity centrality measure ranges from 1 to 5, with a mean of 4.2 and a standard deviation of 0.77. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.51 in the Supplementary Information presents the coefficients for all control variables.

Table 8 presents the results of regressions that interact treatment with moral identity centrality. We find significantly higher treatment effects for our vague allegation outcome among respondents with higher moral identity centrality, consistent with the idea that morally-motivated skepticism drives backlash effects. We observe a directionally consistent interaction term for the antisemitism as a problem outcome, although it is not statistically significant. However, because treatment effects are smaller on this outcome in general, the interaction effect may simply be more difficult to detect.

The evidence is thus suggestive that morally-motivated skepticism drives the core of our backlash

effects, but we lack a design that allows us to convincingly rule out unmotivated skepticism as an alternative mechanism. Future research should develop strategies for further disentangling these phenomena.

6.2 In-Group Defense

We now turn from moral self-defense to the other potential morally-mediated explanation: in-group defense. Perhaps the clearest implication of the in-group defense explanation is that the pro-Palestinian group must be a true social identity for its members. In a pilot survey fielded on Prolific in April–May 2025 (N = 600), we find no evidence that identification with the pro-Palestinian movement rises to the level of a bona fide social identity like race or gender or religion.²² Thus, we believe that our results are not a result of in-group defense.

Furthermore, we present a pre-registered test of the social identity mechanism. If in-group defense drives defensive reactions to allegations of antisemitism, effects should be concentrated among respondents who identify with the movement, rather than respondents who merely passively support pro-Palestine protests. We ask respondents to what extent they agree with the statement, “I identify with the pro-Palestine movement,” a validated measure of social identity group membership (Postmes et al. 2013). In Table 9, we subset our sample into only protest supporters, and interact treatment effects with identification with the Pro-Palestine movement. We do not find clear evidence that identifiers experience significantly stronger backlash than protest supporters who do not identify with the movement. We also continue to observe a significant treatment effect among non-identifier supporters. Because social identity motivations cannot be active for these respondents, backlash cannot depend *solely* on skepticism motivated by in-group defense.

²²This pilot study employed the scale developed by Ashmore et al. (2004), which is widely used and a well-validated battery for measuring and assessing social identity. As shown in Supplemental Information A.5, the correlations between scale factors in the scale we use to measure social identity are substantially lower than we would expect for a true social identity. Principal Component Analysis also suggests that the factors do not load cleanly onto the dimensions we typically see from social identities.

Table 9: Mechanism Test: Pro-Palestine Social Movement Identity

	Vague Future Allegation	Severity of Problem (0-10 Scale)
	Pro-Palestine Protest Supporters	Pro-Palestine Protest Supporters
Treatment	-0.125* (0.051)	-0.251 (0.128)
Pro-Palestine Movement Identification (Standardized)	0.022 (0.037)	0.157 (0.094)
Treatment $\times$ Movement Identification	-0.073 (0.051)	-0.092 (0.128)
Constant	3.353*** (0.139)	5.008*** (0.348)
Observations	2,318	2,320
R^2	0.039	0.071

Note: Treatment effects are derived from OLS models that also control for respondent gender, age, religion, race/ethnicity, partisan identification, self-reported ideology, and education. The dependent variable in the first column is measured on a 1 to 5 scale, with higher values indicating a respondent more strongly believes that the allegation described an actually anti-semitic incident. The dependent variable in the second column ranges from 0 to 10, with higher values indicating that the respondent believes antisemitism is more of a societal problem. Pro-Palestine movement identification is standardized, such that the coefficients reflect the predicted effect of a one standard deviation increase. The original movement identification measure ranges from 1 to 7, with a mean of 3.66 and a standard deviation of 1.72. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Table A.51 in the Supplementary Information present the coefficients for all control variables.

7 Alternative Explanations

7.1 Cognitive Dissonance

One alternative explanation for our results is that survey respondents do not engage in a Bayesian updating process, but instead are merely attempting to avoid experiencing cognitive dissonance (the mental discomfort that comes with holding two beliefs, attitudes, or behaviors that contradict one another) (Festinger 1954). Respondents might view different allegations of antisemitism as heavily linked, particularly if they imagine that these allegations come from the same source. Berinsky and Mendelberg (2005) show that people subconsciously connect different stereotypes about Jewish individuals, even when they consciously accept some stereotypes while rejecting others. Analogously, respondents might subconsciously link distinct *allegations* of antisemitism, even when they can rationally distinguish between them. This linking might still induce cognitive dissonance, making it difficult to accept one accusation while rejecting another. “Why would an organization issue some true but some false accusations?”, a respondent might think.

If respondents do feel cognitive dissonance when rejecting one accusation but accepting another, this could fuel a similar backlash effect to what we observe. Respondents who reject the first treatment accusation could seek to minimize that cognitive dissonance by rejecting subsequent accusations as well. This response pattern would produce a backlash effect, even if respondents do not engage in any kind of informational updating process about the content of the subsequent accusations.

We offer three reasons to believe that cognitive dissonance does not constitute a plausible alternative explanation of our results. First, it is not obvious that respondents should or would feel cognitive dissonance rating allegations differently. It is likely that most people can rationalize differences across accusations. Empirically, this appears to be the case. Across the four allegations in control, we find that 65.26% of respondents do not rate all allegations as equally plausible. This result casts doubt on the notion that respondents feel highly constrained across questions.

Second, we find no evidence that people's preference for consistency, a validated measure of the degree to which people experience cognitive dissonance (Nolan and Nail 2014; Nail et al. 2001; Cialdini et al. 1995), influences treatment effects (see Supplemental Information A.3).

Finally, the pattern of results in Tables 1 and 2 are more consistent with Bayesian updating than cognitive dissonance. As we argue above, the informational mechanism leads us to expect the strongest treatment effects for vague subsequent allegations, because people must rely more on their priors about antisemitism accusations in general when they lack specific details. Cognitive dissonance, on the other hand, should be strongest when two accusations are most similar; we would thus expect the strongest effects on the left-wing accusation, which is most similar to our treatment allegation. As Tables 1 and 2 show, our results are more consistent with the Bayesian account.

7.2 Partisan Cheerleading

Another alternative explanation is that observed backlash is actually a product of partisan cheerleading rather than genuine opinion change. Partisan cheerleading, described by Bullock et al. (2015), refers to the tendency of individuals to answer questions in a way that makes their "team"

look good, despite the fact that they are aware their answers are incorrect, or are aware that they do not know the correct answer. Bullock et al. (2015) detail this cheerleading on behalf of one's political party, but people might feel similarly attached to pro-Palestinian activists as members of their "team." If so, they may pretend to update their beliefs on the severity of antisemitism and the truth of future allegations even when their beliefs do not change, because they believe it makes their "team" look better if antisemitism is not as much of a problem. If this cheerleading is occurring, then we would observe a backlash, but it would be a mirage driven by pretense rather than genuine shifting beliefs.

We attempt to address this possibility following the procedures laid out by Bullock et al. (2015). Bullock and co-authors provide respondents incentives to provide the correct answer to factual questions, or to admit that they do not know the answer. We do the same by presenting half of our respondents with an unincentivized factual question about the number of antisemitic hate crimes recorded by the FBI in 2023, while the other half of respondents received an incentive to answer this question correctly. We present these results in Table A.4 in the Supplemental Information. Because we observe null effects on the hate crimes question even in the unincentivized condition, however, we cannot assess whether monetary incentives reduce the backlash effect.

We nevertheless believe our hate crimes question suggests that partisan cheerleading is an unlikely explanation in this case—if the backlash we document amounts to partisan cheerleading, then the partisan cheerleaders should use the FBI hate crimes question as an opportunity to cheerlead, showing off their devaluation of antisemitic hate crimes.²³

A second reason to doubt the partisan cheerleading account is that partisan cheerleading should be most prevalent among movement identifiers, rather than just movement supporters. We therefore interpret the evidence against social identity group-protective motivations as evidence against partisan cheerleading as well.

²³ On the contrary, we do not believe the null results are particularly problematic for a genuine backlash. Respondents' guesses about hate crimes are extremely noisy, making it difficult for us to observe a treatment effect with our level of power. If respondents are partisan cheerleading, however, they should no longer be anchored to their noisy pre-treatment guesses, and should instead opt to signal their antipathy toward accusations of antisemitism by picking the lowest or near-lowest number of hate crimes.

8 Discussion

How do people respond to expanding definitions of prejudice? Although our study cannot answer this question in full, we focus on one portion of the answer: the backlash effect. In this experiment, we find that exposure to contentious allegations labeling criticism of Israel as antisemitism significantly reduces respondents' belief in subsequent, unrelated allegations of antisemitism and decreases their perception of antisemitism as a serious societal problem.

This backlash may not be driven by skeptics as a whole but by a distinct, yet fairly substantial, subgroup: those who support pro-Palestinian protests. Mechanism tests suggest that moral self-defense—a motivated reasoning process through which individuals reject threatening information to preserve a positive moral self-image—plausibly plays a significant role in the observed backlash. Respondents who perceive contentious allegations as personally accusatory might be discounting subsequent claims of antisemitism as a means of protecting their moral self-image from future accusations that might also be threatening. While the evidence most strongly supports moral self-defense, our design does not allow us to truly exclude backlash in response to unmotivated skepticism as well.

While it might be tempting to label the phenomenon identified in this paper a “boy-who-cried-wolf effect”, this mechanism differs in a crucial sense from the stylized “boy who cried wolf” fable, which most closely approximates backlash through unmotivated skepticism. Backlash may instead only occur when skepticism is motivated by protection of one’s moral self-image, an element not present in the traditional fable. Future work should focus on uncovering whether or not backlash through unmotivated skepticism ever occurs.

Further, we emphasize that the effects on both main outcomes are consistently negative not only in the hypothesized subgroup of skeptics, but in the full sample as well. Additionally, the effects of treatment may not extend beyond these two main outcomes: we detect no significant treatment effects on beliefs about the prevalence of anti-Jewish hate crimes or on endorsement of antisemitic tropes, and we find mixed evidence on willingness to donate to a charity combatting antisemitism and rank order of antisemitism as a problem.

8.1 Practical Implications of this Research

Our findings also shed light on a recent puzzle. From 2023 to 2024—in the midst of the raging definition debate around antisemitism—Americans became much more likely to describe allegations of antisemitism as tools to delegitimize political opponents rather than genuine descriptors of prejudice (Telhami 2025). In other words, public trust in claims of antisemitism appears to be falling. Contrary to these public perceptions, recent work shows that antisemitism is a growing issue in the United States. Research has shown that antisemitic attitudes increased after October 7th, 2023 (Kopstein et al. 2024; Pape 2024), and hate crimes against Jewish people have historically spiked in the wake of Israeli military action (Jacobs et al. 2011; Feinberg 2020; Vergani et al. 2022; Christensen and Enlund 2021). Jewish Americans today represent not just the plurality, but the majority, of victims of religiously-motivated hate crimes (U.S. Department of Justice 2025). So why is trust in antisemitism allegations falling while antisemitism is, in fact, rising? The backlash effect we identify may provide an answer.

The rise of antisemitism makes it particularly consequential that efforts to counter it be both principled and empirically informed. Recent debates among universities, NGOs, and the U.S. Congress have centered on whether to adopt expansive definitions of antisemitism that include criticism of Israel or Zionism (Watanabe 2024; Congress.gov 2025). While activists typically appeal to normative arguments for and against adopting an expansive definition of antisemitism, our findings are the first to shed light on the empirical consequences of this decision. They suggest that labeling events such as pro-Palestinian protests as antisemitic may undermine public concern about antisemitism, contrary to activists' and organizations' stated goals. We caution practitioners to carefully consider the implications of these findings before embracing broad definitions of antisemitism and publicly issuing contentious allegations as institutional charges of prejudice.

As political scientists, we do not make normative judgments about whether expanding the definition of prejudice is morally right or wrong. Organizations that define prejudice broadly may view this approach as ethically necessary despite potential backlash. Our aim is to inform, not prescribe, these actors' strategic decisions. From an instrumental perspective, however, the evidence

indicates that such messaging reduces, rather than increases, public concern about prejudice. Whether this cost outweighs perceived moral imperatives is a question for practitioners themselves.

Two caveats apply to this conclusion. First, some of the hypothesized longer-term or cumulative effects of concept creep, both detrimental—such as a decreased sense of agency among those who are newly identified as victims of prejudice (Haslam 2016)—and beneficial—such as modifying social norms in a constructive manner (Haslam et al. 2020)—cannot be measured in a survey experiment such as this one. Second, as discussed in more detail in Supplemental Information A.9, the negative main effects in the full sample may reflect the fact that skeptics outnumber believers in the population. However, because one’s status as a skeptic or believer is allegation-specific, and because allegations of antisemitism vary in their perceived credibility, our evidence does not necessarily indicate that the negative full-sample ATEs extend to all controversial allegations of antisemitism, only to allegations similar to the one used in our treatment.

8.2 Beyond Antisemitism

While this paper focuses on antisemitism as a test case, we see no reason why this backlash would be restricted to only one form of prejudice. We suspect that this effect may be present in various contexts like racism, sexism, or transphobia. Our general hypothesis is that when one hears an accusation of prejudice which they perceive as unfounded, especially if they feel the accusation targets them, they become less likely to believe other accusations of the same type and less concerned about the form of prejudice generally. As suggested by Dakin et al. (2023), this hypothesis could extend beyond prejudice altogether, to other domains where repeated overreach diminishes credibility; for example, when political actors apply moralized labels (e.g., “fascist”, “genocidal”, or “communist”) to contested cases. We view this paper as only the tip of the iceberg of a larger research agenda which seeks to test this general hypothesis across a variety of contexts.

We also posit an intriguing—albeit speculative—implication of our findings which we suggest future research examine. According to the logic of our theory, concept creep should affect not just perceptions of the severity of prejudice, but also candidate evaluations for politicians accused

of prejudice. Mendelberg (2001) laid out the “norm of racial equality”, outlining that Americans tend to punish overtly racist politicians. Politicians can avoid punishment through subtle racial appeals, and suffer in the polls when the appeals are called out as racist, such as in the Willie Horton episode of 1988 (Mendelberg 2001). But an implication of our backlash hypothesis is that exposure to contentious allegations of prejudice should weaken this punishment, allowing overtly racist candidates, or candidates accused of racism, to more easily win office. Although our study does not test whether this punitive weakening has occurred among the American electorate, we mark this as a question of interest for future research.

Finally, we believe this research agenda may provide crucial insights for organizations seeking to reduce prejudice. How one talks about prejudice may have an enormous impact on the effectiveness of prejudice reduction messaging, and the documented backlash effect may be a key piece in explaining why. We hope that our research may illuminate which types of prejudice reduction messaging are productive and which are counterproductive to the important causes they seek to advance.

References

- Allport, Gordon W. 1954. *The Nature of Prejudice*. Reading, MA: Addison-Wesley.
- Anti-Defamation League. 2024a. “Audit of Antisemitic Incidents 2024 | ADL”.
- Anti-Defamation League. 2024b. “HEAT Map™ | ADL”.
- Antoniou, Georgios, Elias Dinas, and Spyros Kosmidis. 2020. “Collective Victimhood and Social Prejudice: A Post-Holocaust Theory of Anti-Semitism”. *Political Psychology* 41 (5): 861–886.
- Aquino, Karl and Americus II Reed. 2002. “The self-importance of moral identity”. *Journal of Personality and Social Psychology* 83 (6): 1423–1440.
- Ashmore, Richard D., Kay Deaux, and Tracy McLaughlin-Volpe. 2004. “An Organizing Framework for Collective Identity: Articulation and Significance of Multidimensionality”. *Psychological Bulletin* 130 (1): 80–114.
- Balcetis, Emily and David Dunning. 2010. “Wishful seeing: Motivational influences on visual perception of the physical environment”. In *Social Psychology of Visual Perception*, 77–101. Psychology Press.
- Bassi, Camila. 2023. “Geography, antisemitism and Zionism”. *Geography (Sheffield, England)* 108 (3): 130–133.
- Bayes, Robin and James N Druckman. 2021. “Motivated reasoning and climate change”. *Current Opinion in Behavioral Sciences* 42 : 27–35.
- Beller, Steven. 2015. *Antisemitism: A Very Short Introduction* (Second Edition, New to this Edition:, Second Edition, New to this Edition: ed.). Oxford University Press.
- Berinsky, Adam J. and Tali Mendelberg. 2005. “The indirect effects of discredited stereotypes in judgments of jewish leaders”. *American Journal of Political Science* 49 (4): 845–864.

- Berkman, Matthew. 2021. "Antisemitism, Anti-Zionism, and the American Racial Order: Revisiting the American Council for Judaism in the Twenty-First Century". *American Jewish History* 105 (1): 127–155.
- Bilewicz, Michal. 2022. "Explaining the Jew-Hatred: The structure and psychological antecedents of antisemitic beliefs". In *The Routledge International Handbook of Discrimination, Prejudice and Stereotyping*. Routledge.
- Bishin, Benjamin G, Thomas J Hayes, Matthew B Incantalupo, and Charles Anthony Smith. 2016. "Opinion backlash and public attitudes: Are political advances in gay rights counterproductive?". *American Journal of Political Science* 60 (3): 625–648.
- Blum, Lawrence. 2002. " *I'm Not a Racist, But...*" *The Moral Quandary of Race*. Cornell University Press.
- Bolsen, Toby and Risa Palm. 2019. "Motivated Reasoning and Political Decision Making". In *Oxford Research Encyclopedia of Politics*. Oxford University Press.
- Bullock, John G. 2009. "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion". *The Journal of Politics* 71 (3): 1109–1124.
- Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. 2015. "Partisan Bias in Factual Beliefs about Politics". *Quarterly Journal of Political Science* 10 (4): 519–578.
- Carlson, Erik. 2025. "'never again is now:' use of cultural frames in a social movement organization". *Transdisciplinary Journal of Management*.
- Cascardi, Michele and Cathy Brown. 2016. "Concept creep or meaningful expansion? response to haslam". *Psychological Inquiry* 27 (1): 24–28.
- Christensen, Love and Jakob Enlund. 2021. "Echoes of Violent Conflict: The Effect of the Israeli-Palestinian Conflict on Hate Crimes in the U.S.". *Working Papers in Economics*.

- Cialdini, Robert B., Melanie R. Trost, and Jason T. Newsom. 1995. "Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications". *Journal of Personality and Social Psychology* 69 (2): 318–328.
- Cohen, Florette, Lee Jussim, Kent D. Harber, and Gautam Bhasin. 2009. "Modern anti-Semitism and anti-Israeli attitudes". *Journal of Personality and Social Psychology* 97 (2): 290–306.
- Collier, David and Steven Levitsky. 1996. "Democracy with adjectives: Conceptual innovation in comparative research". *World Politics* 49 (3): 430–451.
- Collier, David and James E. Mahon, Jr. 1993. "Conceptual "stretching" revisited: Adapting categories in comparative analysis". *American Political Science Review*, 87 (4): 845–855.
- Concerned Jewish Faculty and Staff. 2024. "Statement from Concerned Jewish Faculty Against Antisemitism".
- Congress.gov. 2025. "S.558 - 119th Congress (2025-2026): Antisemitism Awareness Act of 2025".
- Coppock, Alexander. 2023. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.
- Dakin, Brodie C, Melanie J McGrath, Joshua J Rhee, and Nick Haslam. 2023. "Broadened concepts of harm appear less serious". *Social Psychological and Personality Science* 14 (1): 72–83.
- Ding, Peng and Jiannan Lu. 2017. "Principal stratification analysis using principal scores". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79 (3): 757–777.
- Dinnerstein, Leonard. 1995. *Antisemitism in America*. Oxford University Press.
- Ditto, Peter H. and David F. Lopez. 1992. "Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions". *Journal of Personality and Social Psychology* 63 (4): 568–584.

- Ditto, Peter H., David A. Pizarro, and David Tannenbaum. 2009. "Chapter 10 Motivated Moral Reasoning". In *Psychology of Learning and Motivation*, Volume 50, 307–338. Academic Press.
- Dixon, John, Mark Levine, Steve Reicher, and Kevin Durrheim. 2012. "Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution?". *Behavioral and brain sciences* 35 (6): 411–425.
- Dovidio, John F. and Samuel L. Gaertner. 1986. *Prejudice, discrimination, and racism*. Academic Press.
- Feinberg, Ayal. 2020. "Explaining Ethnoreligious Minority Targeting: Variation in U.S. Anti-Semitic Incidents". *Perspectives on Politics* 18 (3): 770–787.
- Festinger, Leon. 1954. "A theory of social comparison processes". *Human Relations* 7 (2): 117–140.
- Foxman, Abraham H. 2003. *Never Again? The Threat of the New Anti-Semitism*. HarperSanFrancisco.
- Frangakis, Constantine E and Donald B Rubin. 2002. "Principal stratification in causal inference". *Biometrics* 58 (1): 21–29.
- Frindte, Wolfgang, Susan Wettig, and Dorit Wammetsberger. 2005. "Old and New Anti-Semitic Attitudes in the Context of Authoritarianism and Social Dominance Orientation—Two Studies in Germany". *Peace and Conflict: Journal of Peace Psychology* 11 (3): 239–266.
- Gaines, Brian J., James H. Kuklinski, Paul J. Quirk, Buddy Peyton, and Jay Verkuilen. 2007. "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq". *The Journal of Politics* 69 (4): 957–974.
- Gerber, Alan and Donald Green. 1999. "Misperceptions about Perceptual Bias". *Annual Review of Political Science* 2 (Volume 2, 1999): 189–210.
- Gerring, John. 1999. "What makes a concept good? a criterial framework for understanding concept formation in the social sciences". *Polity* 31 (3): 357–393.

- Glick, Peter and Susan T. Fiske. 1996. "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism". *Journal of Personality and Social Psychology* 70 (3): 491–512.
- Hartman, Erin and Melody Huang. 2024. "Improving precision through design and analysis in experiments with noncompliance". *Political Science Research and Methods* 12 (3): 557–572.
- Haslam, Nick. 2016. "Concept creep: Psychology's expanding concepts of harm and pathology". *Psychological Inquiry* 27 (1): 1–17.
- Haslam, Nick, Brodie C Dakin, Fabian Fabiano, Melanie J McGrath, Joshua Rhee, Ekaterina Vylomova, Morgan Weaving, and Melissa A Wheeler. 2020. "Harm inflation: Making sense of concept creep". *European Review of Social Psychology* 31 (1): 254–286.
- Heere, Bob and Jeffrey D. James. 2007. "Sports teams and their communities: Examining the influence of external group identities on team identity". *Journal of Sport Management* 21 (3): 319–337.
- Herek, Gregory M. 2000. "The psychology of sexual prejudice". *Current Directions in Psychological Science* 9 (1): 19–22.
- Hersh, Eitan and Laura Royden. 2022. "Antisemitic Attitudes Across the Ideological Spectrum". *Political Research Quarterly* 76 (2): 697–711.
- Howell, Jennifer L., Redford Liz Pogge Gabrielle and Kate A. Ratliff. 2017. "Defensive responding to iat feedback". *Social Cognition* 35 (5): 520–562.
- IHRA. 2016. "Working Definition of Antisemitism".
- Jacobs, Dirk, Yoann Veny, Louise Callier, Barbara Herman, and Aurélie Descamps. 2011. "The impact of the conflict in Gaza on antisemitism in Belgium". *Patterns of Prejudice* 45 (4): 341–360.

- Jordan, Jennifer, Elizabeth Mullen, and J. Keith Murnighan. 2011. "Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior". *Personality and Social Psychology Bulletin* 37 (5): 701–713.
- Judaken, Jonathan. 2008. "So what's new? rethinking the 'new antisemitism' in a global age". *Patterns of Prejudice* 42 (4/5): 531–560.
- Kahan, Dan M. 2013. "Ideology, motivated reasoning, and cognitive reflection". *Judgment and Decision Making* 8 (4): 407–424.
- Kaleem, Jaweed. 2025. "Judge blocks Trump administration push to fine UCLA \$1.2 billion for alleged antisemitism". *The Los Angeles Times*.
- Kaplan, Edward H. and Charles A. Small. 2006. "Anti-Israel Sentiment Predicts Anti-Semitism in Europe". *The Journal of Conflict Resolution* 50 (4): 548–561.
- Kendi, Ibram X. 2019. *How to be an Antiracist*. One World.
- Kinder, Donald R. and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. University of Chicago Press.
- Kinder, Donald R. and David O. Sears. 1981. "Prejudice and politics: Symbolic racism versus racial threats to the good life". *Journal of Personality and Social Psychology* 40 (3): 414–431.
- King, Ryan D. and Melissa F. Weiner. 2007. "Group Position, Collective Threat, and American Anti-Semitism". *Social Problems* 54 (1): 47–77.
- Klug, Brian. 2003. "The collective jew: Israel and the new antisemitism". *Patterns of Prejudice* 37 (2): 117–138.
- Kofta, Mirosław, Wiktor Soral, and Michał Bilewicz. 2020. "What Breeds Conspiracy Anti-semitism? The Role of Political Uncontrollability and Uncertainty in the Belief in Jewish Conspiracy". *Journal of Personality and Social Psychology: Attitudes and Social Cognition* 118 (5): 900–918.

- Kopstein, Jeffrey, Ana Schugurensky, and Rachel Shenhav-Goldberg. 2024. “Attitudes Toward Jews and Israel on California Campuses: Results from a New Survey | ADL”. Technical report, Anti-Defamation League (ADL).
- Kunda, Ziva. 1990. “The case for motivated reasoning”. *Psychological Bulletin* 108 (3): 480–498.
- Lodge, Milton and Charles S Taber. 2005. “The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis”. *Political psychology* 26 (3): 455–482.
- Lofaro, Nicole, Louis H. Irving, and Kate A. Ratliff. 2024. “Defensiveness Toward IAT Feedback Predicts Willingness to Engage in Anti-Bias Behaviors”. *Personality and Social Psychology Bulletin*: 01461672231219948.
- Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence”. *Journal of Personality and Social Psychology* 37 (11): 2098–2109.
- McConahay, John B. and Joseph C. Hough Jr. 1976. “Symbolic racism”. *Journal of Social Issues* 32 (2): 23–45.
- Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.
- Meuleman, Bart, Koen Abts, Koen Sloopmaeckers, and Cecil Meeusen. 2019. “Differentiated Threat and the Genesis of Prejudice: Group-Specific Antecedents of Homonegativity, Islamophobia, Anti-Semitism, and Anti-Immigrant Attitudes”. *Social Problems* 66 (2): 222–244.
- Miles, Robert. 1989. *Racism*. Routledge.
- Monin, Benoît and Alexander H Jordan. 2009. “The dynamic moral self: A social psychological perspective”. *Personality, identity, and character: Explorations in moral psychology*: 341–354.

- Mulder, Laetitia B. and Karl Aquino. 2013. "The role of moral identity in the aftermath of dishonesty". *Organizational Behavior and Human Decision Processes* 121 (2): 219–230.
- Nail, Paul R, Jayme S Correll, Chad E Drake, Shyla B Glenn, Gina M Scott, and Careylyn Stuckey. 2001. "A validation study of the preference for consistency scale". *Personality and Individual Differences* 31 (7): 1193–1202.
- Nexus Project. 2024. "The Nexus Document".
- Nolan, Jon and Paul Nail. 2014. "Further Evidence That Individuals With a High Preference for Consistency Are More Susceptible to Cognitive Dissonance". *Psi Chi Journal of Psychological Research* 19 (4): 214–219.
- Pape, Robert A. 2024. "Understanding Campus Fears After October 7 and How to Reduce Them". Technical report, Chicago Project on Security & Threats.
- Pierce, C. 1974. "Psychiatric problems of the black minority". *American Handbook of Psychiatry* 2 : 512–523.
- Postmes, Tom, S. Alexander Haslam, and Lise Jans. 2013. "A single-item measure of social identification: Reliability, validity, and utility". *British Journal of Social Psychology* 52 (4): 597–617.
- Rosenfeld, Alvin H. 2019. "6. Leaving the Post-Holocaust Period: The Effects of Anti-Israel Attitudes on Perceptions of the Holocaust". In *Anti-Zionism and Antisemitism: The Dynamics of Delegitimization*, 158–174. Indiana University Press.
- Sartori, Giovanni. 1970. "Concept misinformation in comparative politics". *American Political Science Review*, 64 (4): 1033–1053.
- Snow, David and Catherine Corrigan-Brown. 2005. "Falling on deaf ears: Confronting the prospect of nonresonant frames". In *Rhyming Hope and History: Activism and Social Movement Scholarship*. University of Minnesota Press.

- Solomos, John. 2017. "Are racists crazy? How prejudice, racism, and antisemitism became makers of insanity". *Ethnic and Racial Studies* 41 (3): 606–608.
- Solorzano, Daniel, Miguel Ceja, and Tara Yosso. 2000. "Critical race theory, racial microaggressions, and campus racial climate: The experiences of african american college students". *The Journal of Negro Education* 69 (1/2): 60–73.
- Stanislowski, Michael. 2017. *Zionism: A Very Short Introduction*, Volume 507. Oxford University Press.
- Steele, Claude M. 1988. "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self". In *Advances in Experimental Social Psychology*, ed. Leonard Berkowitz, Volume 21, 261–302. Academic Press.
- Stempel, Jonathan. 2025. "Harvard settles Jewish student's lawsuit over alleged antisemitism". *Reuters*.
- Sue, Derald Wing, Christina M. Capodilupo, Gina C. Torino, Jennifer M. Bucceri, Aisha M. B. Holder, Kevin L. Nadal, and Marta Esquilin. 2007. "Racial microaggressions in everyday life: Implications for clinical practice". *The American Psychologist* 62 (4): 271–286.
- Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs". *American Journal of Political Science* 50 (3): 755–769.
- Tajfel, Henri and John Turner. 1979. "An integrative theory of intergroup conflict". In *The Social Psychology of Intergroup Relations*, eds. William G Austin and Stephen Worchel, 94–109. Brooks/Cole Publishing Company.
- Telhami, Shibley. 2025. "Poll analysis: Trump's approach to universities, Israel-Palestine, and antisemitism".
- Tetlock, Philip E. 1994. "Political psychology or politicized psychology: Is the road to scientific hell paved with good moral intentions?". *Political Psychology* 15 (3): 509–529.

U.S. Department of Justice. 2025. “2023 FBI Hate Crimes Statistics”.

Vergani, Matteo, Dan Goodhardt, Rouven Link, Amy Adamczyk, Joshua D. Freilich, and Steven Chermak. 2022. “When and How Does Anti-Semitism Occur? The Different Trigger Mechanisms Associated with Different Types of Criminal and Non-Criminal Hate Incidents”. *Deviant Behavior*.

Vitriol, Joseph A. and Gordon B Moskowitz. 2021. “Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change”. *Journal of Experimental Social Psychology* 96 : 104–165.

Watanabe, Teresa. 2024. “UC Berkeley to expand antisemitism education to all new students”. *The Los Angeles Times*.

Waxman, Dov, David Schraub, and Adam Hosein. 2022. “Arguing about antisemitism: Why we disagree about antisemitism, and what we can do about it”. *Ethnic and Racial Studies* 45 (9): 1803–1824.

Weinberg, George. 1972. *Society and the healthy homosexual*. St. Martin’s Press.

Westen, Drew, Pavel S Blagov, Keith Harenski, Clint Kilts, and Stephan Hamann. 2006. “Neural bases of motivated reasoning: An fmri study of emotional constraints on partisan political judgment in the 2004 us presidential election”. *Journal of cognitive neuroscience* 18 (11): 1947–1958.

Wright, Graham, Shahar Hecht, Sasha Volodarsky, and Leonard Saxe. 2024. “Antisemitism on Campus: Understanding Hostility to Jews and Israel”. Technical report, Brandeis University.

Yazdiha, Hajar. 2017. *The Legacies of the Civil Rights Movement and the Paradoxical Politics of Inclusion: Collective Memory in Contentious Politics*. Phd dissertation, University of North Carolina at Chapel Hill.