Blatant Dehumanization in the Mind's Eye:

Prevalent Even among Those Who Explicitly Reject It?

Christopher D. Petsko

Duke University

Ryan F. Lei

Haverford College

Jonas R. Kunst

University of Oslo

Emile Bruneau

University of Pennsylvania & Beyond Conflict Innovation Lab

Nour Kteily

Northwestern University

This manuscript has been accepted for publication at *Journal of Experimental Psychology: General*. Because this manuscript has not yet been processed by the journal's editorial office, this version of the article may differ slightly from the final version that is published in *JEP:G*. Prior to publication, this paper can be cited as follows:

Petsko, C. D., Lei, R. F., Kunst, J. R., Bruneau, E., & Kteily, N. (in press). Blatant dehumanization in the mind's eye: Prevalent even among those who explicitly reject it? *Journal of Experimental Psychology: General.*

Correspondence concerning this article can be sent either to Christopher Petsko (christopher.petsko@duke.edu) or to Nour Kteily (n-kteily@kellogg.northwestern.edu).

Abstract

Research suggests that some people, particularly those on the political right, have a tendency to blatantly dehumanize low-status groups. However, these findings have largely relied on self-report measures, which are notoriously subject to social desirability concerns. To better understand just how widely blatant forms of intergroup dehumanization might extend, the present paper leverages an unobtrusive, data-driven perceptual task to examine how U.S. respondents mentally represent 'Americans' vs. 'Arabs' (a low-status group in the U.S. that is often explicitly targeted with blatant dehumanization). Data from two reverse-correlation experiments (original $N = 108$; pre-registered replication $N = 336$) and seven rating studies ($N = 2{,}301$) suggest that U.S. respondents' mental representations of Arabs are significantly more dehumanizing than their representations of Americans. Furthermore, analyses indicate that this phenomenon is not reducible to a general tendency for our sample to mentally represent Arabs more negatively than Americans. Finally, these findings reveal that blatantly dehumanizing representations of Arabs can be *just as prevalent* among individuals exhibiting low levels of explicit dehumanization (e.g., liberals) as among individuals exhibiting high levels of explicit dehumanization (e.g., conservatives)—a phenomenon into which exploratory analyses suggest liberals may have only limited awareness. Taken together, these results suggest that blatant dehumanization may be more widespread than previously recognized, and that it can persist even in the minds of those who explicitly reject it.

Keywords: dehumanization, mental representations, reverse correlation, prejudice, intergroup relations

Blatant Dehumanization in the Mind's Eye: Prevalent Even among Those Who Explicitly Reject It?

Blatant forms of dehumanization are often assumed to be a relic of our troubled past. However, recent research illustrates that blatant forms of dehumanization persist, with perceivers around the world consistently likening certain (typically low-status) groups of people to non-human animals (e.g., Jackson & Gaertner, 2010; for reviews, see Bar-Tal, 1989; Kteily & Bruneau, 2017b). For example, individuals in the United States—the vast majority of whom are not Arab—tend to rate Arabs as 'less evolved' than other groups of people. This phenomenon is consequential, as the tendency to blatantly dehumanize low-status groups of people (e.g., Arabs, Mexican immigrants) in this way predicts support for behaviors and policies that disadvantage them. For example, blatant dehumanization of Arabs in the United States predicts support for restricting Arab immigration to the U.S., and it also predicts support for the use of extreme counterterrorism tactics on Arab people (including torture). Importantly, blatant dehumanization continues to predict these outcomes even when controlling for perceivers' general dislike of dehumanized groups (Kteily, Bruneau, Waytz, & Cotterill, 2015) as well as for their tendency to underestimate these groups' mental and emotional capacities (i.e., subtler forms of dehumanization; e.g., Leyens et al., 2000; Waytz, Gray, Epley, & Wegner, 2010). Put simply, the existing evidence suggests that blatant dehumanization is pervasive, that it is consequential for those it targets, and that it is non-reducible to 'mere' prejudice or subtler forms of dehumanization (Bruneau, Jacoby, Kteily, & Saxe, 2018).

Still, the existing research on blatant dehumanization provides only a limited understanding of its internal workings (but see Goff, Eberhardt, Williams, & Jackson, 2008, who

focus specifically on the automatic association between "Black" and "ape"). When participants think about a blatantly dehumanized group—like Arabs, in the U.S. context—what do they imagine in their mind's eye? And how does this relate to any dehumanization that they outwardly express or, perhaps, withhold? Given the importance of dehumanization to real-world social issues and intergroup conflict (Bruneau & Kteily, 2017a; Bruneau, Kteily, & Laustsen, 2018; Goff et al., 2008; Goff, Jackson, Di Leone, Culotta, & DiTomasso, 2014; Viki, Osgood, & Phillips, 2013), better understanding how dehumanization manifests in the mind is of key practical interest. And from a theoretical perspective, understanding how people who self-report no blatant dehumanization of a target group mentally conceive of that group is particularly important. Do their mental representations match the full humanness that they explicitly attribute to these groups, or do they harbor recognizably dehumanizing representations of these groups despite outwardly rejecting any denial of their humanity?[1]

To investigate these questions, we leveraged a novel, unobtrusive methodology—the reverse-correlation technique (Dotsch & Todorov; 2012; Mangini & Biederman, 2004)—to create composite images that index perceivers' mental representations of Arabs, a low-status group that is often the target of explicit dehumanization in the U.S. In this task, perceivers are instructed to view pairs of black-and-white facial images that are overlaid with random visual noise. Their task is to select the face in each pair that looks most similar to a given target group (e.g., Arabs). Perceivers do this across hundreds of trials, allowing researchers to create

[1] We note that we focus in this paper specifically on blatant dehumanization (i.e., representations of a group that are directly interpretable as reflecting a perceived lack of full humanity, such as linking a group with non-human metaphors or clearly animalistic traits such as 'savage'). Other research has focused on more subtle forms of dehumanization, in which dehumanization can be indirectly inferred from perceivers withholding some capacity associated with humanity (e.g., the capacity to fully experience complex emotions or human-related traits) from a target. We note that some of the work on more subtle forms of dehumanization has considered how it might manifest similarly or differently at each of the explicit and implicit levels (e.g., Loughnan & Haslam, 2007; Loughnan, Haslam, & Kashima, 2009; Saminaden, Loughnan, & Haslam, 2010), albeit not using the reverse-correlation technique.

aggregated composite images of the faces that perceivers chose during the task. Aggregating perceivers' choices into composite images causes random visual noise across the facial images to cancel itself out; more importantly, it causes *non-random* features of their selections to become accentuated. This process results in composite images that reflect, in principle, how perceivers mentally represent the target group(s) in question. Once created, these composite images can then be rated by naïve participants on any dimension of interest. Here, we use the reverse correlation task to examine whether targets belonging to a low-status group (i.e., Arabs) are indeed mentally represented in ways that are seen by naïve raters as blatantly dehumanizing.

Of note, a handful of existing studies have used reverse correlation to provide evidence for this kind of dehumanization—specifically, for dehumanization that can be visibly detected from the way a group of people is mentally represented (Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2016; Kunst, Kteily, & Thomsen, 2017). However, these studies did not disentangle dehumanization in the mind's eye from its confounding variables, like the fact that outgroup members (vs. ingroup members) tend to be mentally represented with features that are altogether more negative-looking (Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014). In the present analyses, we estimate how dehumanizing representations of Arabs are while controlling for the extent to which Arabs are represented with negative-looking features. In addition, the analyses in this paper control for whether *raters themselves* are explicitly biased against Arabs (that is, whether people who *rate* the mental representations are themselves prejudiced against or blatantly dehumanizing of Arab individuals). This latter control is added to help ensure that any dehumanization that is detected from mental representations is in fact due to characteristics of the mental representations themselves—rather than to, for example, raters perceiving 'Arab'

composite images as Arab-looking and projecting their own anti-Arab prejudices onto those images.

After providing estimates of how much mentally represented dehumanization, if any, remains after partialing out these confounds, we turn to the question of whether dehumanization in the mind's eye varies across levels of explicit (i.e., self-reported) dehumanization. Past research has shown that explicitly expressed attitudes and beliefs about a group can indeed correlate with how that group is mentally represented. For example, the tendency to represent low-status people as looking stereotypically Black is attenuated among those who are lower (vs. higher) in prejudice against the poor (Lei & Bodenhausen, 2017). Similarly, the tendency to represent feminists as masculine-looking is attenuated among those who are lower (vs. higher) in hostile sexism (Gundersen & Kunst, 2018). This suggests that dehumanization in one's mental representations may be attenuated among those who self-report lower levels of blatant dehumanization (although no existing research has examined this possibility). Still, even if it is lower, the *extent* matters. That is, it would have very different implications for the prevalence of blatantly dehumanizing perceptions if those who self-report low blatant dehumanization have humanizing mental representations that look vastly different from those who self-report high blatant dehumanization versus dehumanizing mental representations that are difficult to distinguish from those of high explicit dehumanizers.

After considering the overlap between self-reported dehumanization and dehumanization in one's mental representations, we consider how mental representations of Arabs might vary across the political spectrum. Specifically, we compare how dehumanizing the mental representations of Arabs are among political conservatives and political liberals (who are less likely than conservatives to express blatant dehumanization of low-status groups on self-report

measures; e.g., Kteily et al., 2015). Past research examining affective prejudice suggests that although liberals harbor less prejudice than conservatives on both the implicit and explicit levels, the gap is smaller at the implicit level (Nosek, Banaji, & Jost, 2009). If liberals' *prejudice* against derogated groups can be higher when measured using unobtrusive (vs. self-report) measures, the same may be true of liberals' *dehumanizing representations* of derogated groups. Finally, we conclude with an exploratory investigation of the degree to which individuals have insight into whether their own mental representations are dehumanizing.

Overview

Throughout this manuscript, people who participated in the reverse-correlation task (and whose choices were used to generate the composite images) are referred to as image *generators.* People who rated composite images (e.g., on how dehumanizing they appear) are referred to as image *raters*. This manuscript presents data from two independent samples of generators: (1) a convenience sample acquired from Mechanical Turk (MTurk), and (2) a Qualtrics Panels sample that was designed to approximate the demographic features of the U.S. population in terms of age, gender, race, socioeconomic status, geographic location, and political group membership. Images from the MTurk sample were rated and analyzed first. We then replicated our findings in the much larger and more representative Qualtrics Panels sample (the analyses for which were pre-registered). Because the methods used to generate (and rate) composite images were almost identical across the two samples, we report the methods and procedures for these two samples simultaneously. This paper reports all exclusions and conditions. Data, materials, pre-registration documentation, and analysis scripts for this paper are available on the Open Science Framework website (OSF: https://osf.io/5mwpq).

Image Generation and Image Rating Methods

In the Image Generation Studies, generators from MTurk and Qualtrics Panels, respectively, completed a reverse-correlation procedure in which they were randomly assigned to select either a) which images looked 'more Arab' or b) which images looked 'more American.'[2] In the Image Rating Studies, convenience samples of online participants rated the composite images that resulted from generators' selections.

Image Generation Studies

Participants. Image generators from the MTurk sample (N = 108) were recruited through TurkPrime.com (Litman, Robinson, & Abberbock, 2017), and were selected on the basis of political lean (52.8% scored at a 3 or below on a scale from *0 = extremely liberal* to *10 = extremely conservative*; 47.2% scored at a 7 or above). Image generators from the Qualtrics Panels sample (N = 336) were selected to approximate the distribution of the U.S. population on age, gender, race/ethnicity, education level, income level, geographic location, and political group membership.[3] The MTurk sample had 77 White, 10 Asian, 7 Black, 5 Latinx, 1 Pacific Islander, and 8 race-non-specified participants; 54 male, 46 female, and 8 gender-non-specified participants; and it had ages that ranged from 21 to 72 (M = 37.31, SD = 11.34). The Qualtrics Panels sample had 208 White, 42 Black, 17 Asian, 60 Latinx, 7 other-identified participants; it had 192 female and 144 male participants; and it had ages ranging from 18 to 95 (M = 47.32, SD = 16.68). None of image generators in either sample identified as Arab.

Procedure. Generators in both samples completed a reverse-correlation task in which they viewed 300 pairs of blurry, black-and-white facial images. Their task was to select the face

[2] We do not mean to suggest (or reify) the idea that "Arab" and "American" are mutually exclusive groupings. Rather, we focus our analysis on "Arabs" vs. "Americans" because people in the United States reliably characterize these two groups as seeming differentially humanized on self-report measures (see e.g., Kteily et al., 2015).
[3] Of note, we had originally targeted 400 participants, but Qualtrics Panels exhausted their participant pool while trying to accommodate our sampling requests (given the quotas we targeted to ensure relative representativeness), and we were therefore only able to collect data from a total of 336 participants.

in each pair that looked most typical of the target group. In every pair of faces, one image was created by adding random visual noise to a base image, and the other was created by adding the inverse of that noise to the same base image (see Figure 1). The base image was an averaged, neutrally expressive male face (taken from the AKDEF database; Lundqvist & Litton, 1998). The creation of noise-imbued face pairs from this base image was conducted using the "rcicr" package in R (Dotsch, 2016). By random assignment, generators either answered the question "Who looks more American?" in every pair, or the question "Who looks more Arab?" in every pair. The ordering of face pairs was randomized for each generator.

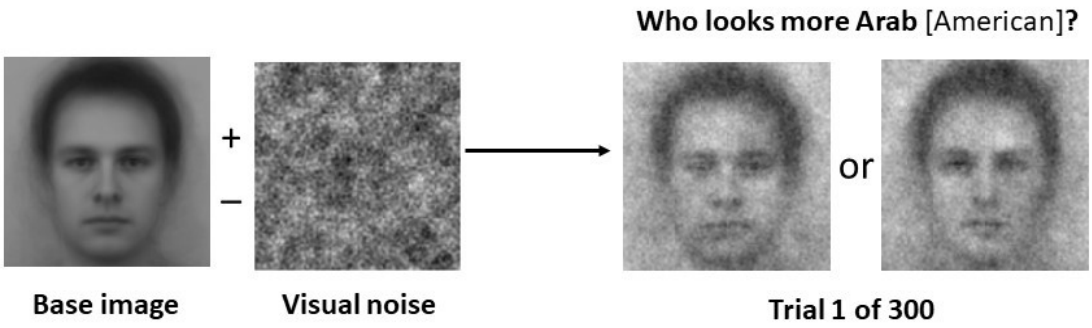

Figure 1. One of 300 possible reverse correlation trials. Each trial is a forced-choice task between two faces: one resulting from adding random visual noise to a base image, the other resulting from subtracting that same visual noise from the same base image.

Generators in both samples then completed three measures of left *vs.* right wing political leanings: a one-item measure of political ideology (an 11-point scale anchored at *0 = extremely liberal* and *10 = extremely conservative*; Kroh, 2007); a measure of social dominance orientation (SDO_7; αs = .89–.96; sample item: "Some groups of people are simply inferior to other groups"; Ho et al., 2015); and a measure of right-wing authoritarianism (RWA). The MTurk sample completed a 20-item version of the RWA measure (α = .97; sample item: "Young people sometimes get rebellious ideas, but as they grow up they ought to get over them and settle down"; Altemeyer, 2007); the Qualtrics Panels sample completed a recently validated 6-item

short version of the RWA scale (α = .68; Bizumic & Duckitt, 2018). Finally, generators completed two self-report measures of blatant dehumanization, summarized below.[4]

Trait dehumanization. Generators indicated the extent to which they thought Arabs and Americans could be characterized by 11 blatantly dehumanizing traits (e.g., "lacking self-restraint, like animals"; "savage, aggressive"; adapted from Bastian, Denson, & Haslam, 2013) on a scale from *1* (*not at all*) to *7* (*very much*). We computed average ratings for each target group (Arabs, αs = .91–.96; Americans, αs = .88–.91). Trait dehumanization was indexed by the extent to which generators used these traits to characterize Arabs more than Americans. On average, generators in both samples exhibited significant levels of blatant dehumanization on this measure (MTurk: M_{diff} = 1.23, β = 0.78, $p < .001$; Qualtrics Panels: M_{diff} = 0.50, β = 0.39, $p < .001$).[5]

Ascent dehumanization. Generators also indicated how evolved they regarded Americans and Arabs (among several distractor groups) to be by rating these groups along a continuum beneath the 'Ascent of (hu)man' diagram (see Kteily et al., 2015). Generators placed each target group on a sliding scale that was anchored at *0* (corresponding to the *least-evolved* scale value) and *100* (the *most-evolved* scale value). Ascent dehumanization was indexed by the extent to which generators regarded Arabs as 'less evolved' than Americans. On average, generators in both samples exhibited significant levels of blatant dehumanization on this measure (MTurk: M_{diff} = 20.97, β = 0.76, $p < .001$; Qualtrics Panels: M_{diff} = 13.41, β = 0.51, $p < .001$).

Image Rating Studies

[4] Generators in the Qualtrics Panels sample also completed several exploratory measures (e.g., how much personal contact they report having with Arabs). We discuss some of these exploratory measures in a later section of the paper, but see OSF for greater detail on what we included and found.

[5] Standardized effect sizes (βs) for experimental designs were always computed by regressing z-standardized outcomes onto orthogonal condition contrasts (which summed to zero and always had a range of one). Under this analytic strategy, βs from experimental studies can be interpreted similarly to Cohen's *d*s.

From the generators' data, we created two classes of images—representing two classes of mental representations—for raters to evaluate in the image rating studies. In some rating studies, raters evaluated individual-level composite images; in others, raters evaluated group-level composite images. By *individual-level composite images*, we refer to images that represent the mental representations of individual generators. By *group-level composite images*, we refer to images that represent the averaged mental representation of a *group* of generators (e.g., all generators in the Arab condition; or the group of generators in the Arab condition who were above the median on SDO). Individual-level composite images were always rated by samples of approximately $n = 30$ people per image; group composite images were always rated by samples of approximately $n = 100$ people per image. Sample sizes for rating studies were determined a priori; implications for statistical power will be discussed below for each analysis, as different rating studies had different experimental designs (see Table S1 for a description of all rating studies, including per-study *N*s and experimental designs).

Of note, creating group-level mental representations at varying levels of a moderator can, at least in some contexts, be statistically problematic—for example, because it can require researchers to conduct median splits on their data (e.g., McClelland, Lynch, Irwin, Spiller, & Fitzsimons, 2015). Nevertheless, we do so here for several reasons. First, group-level composite images collapse across individual images that contain idiosyncratic visual noise patterns (Dotsch & Todorov, 2012). Insofar as visual noise is reduced through the process of aggregation, consensual properties of how a target group is represented can become more clearly accentuated by group composite images than by individual composite images. This is to say that group-level composite images can capture information that is quite literally easier "see" than that captured by more fine-grained individual composite images. Second, the creation of group-level composite

images *as well as* individual-level composite images enables us to examine whether our results hold up to a variety of analytic approaches—for example, to categorical and continuous tests of statistical moderation. Third, the creation of group-level composite images at varying levels of a moderating factor is normative in reverse correlation research (e.g., Gundersen & Kunst, 2018; Lei & Bodenhausen, 2017), thus making our analytic approach and findings comparable to those of other research teams.

Participants. A total of $N = 2{,}403$ MTurk workers were recruited to participate in one of seven image rating studies. Of these, we eliminated $n = 102$ (4.2%) from our final analyses because they did not reply "yes" to the question, "Did you take this survey seriously?"[6] Our final sample of $N = 2{,}301$ raters had 1,686 White, 230 Black, 187 Asian, 124 Latinx, 21 Native American, 50 other, and 3 race-non-specified participants; 1,210 men, 1,071 women, 11 other, and 9 gender-non-specified participants; and it had ages ranging from 18 to 79 ($M = 35.93$, $SD = 10.98$).

Procedure. Regardless of the study in which raters were enrolled, raters were told that the research team was "interested in how people evaluate different kinds of images that vary in their clarity." And that they had "been assigned to evaluate faces that are moderately blurry." Participants who were in group-level image rating studies always saw four photos in a within-person experimental design. Participants who were in an individual-level image rating study always saw an even number of individual composite images of Arabs and Americans, respectively, drawn randomly from the total pool of individual composite images from either the MTurk sample of generators or the Panel sample of generators (see Table S1 for more detail).

[6] The rating study described in Row 2 of Table S1 also excluded all participants who did not complete a second wave of data collection, which is where that particular rating study collected its covariates. These exclusions were pre-registered (see OSF) and they do not change any of this paper's conclusions.

Analytic strategies for rating studies will be discussed as they become relevant to our research questions.

Dependent variables. Regardless of study, raters indicated the extent to which the people represented in the images looked dehumanized on the same two measures as described previously: the 11-item trait dehumanization measure (αs = .87–.94), and the ascent dehumanization measure. Throughout this paper, we collapse these two dependent measures by z-standardizing each of them and then averaging them into a single index of how dehumanizing the people represented in the composite images appear (see Kteily & Bruneau, 2017a, for a similar analytic approach).[7]

Covariates. Regardless of study, raters then indicated the extent to which they felt warm/favorable toward the people represented in each of their assigned images (anchored at *0 = very cold* and *100 = very warm*). This enabled us to estimate how dehumanizing mental representations of Arabs (vs. Americans) appear while covarying out the extent to which generators represent Arabs more negatively than Americans. In addition, most rating studies (exceptions will be noted, where applicable) included two additional covariates: (a) the extent to which raters themselves tend to explicitly dehumanize Arabs relative to Americans (assessed by asking raters to complete the ascent measure); and (b) the extent to which raters themselves are evaluatively prejudiced toward Arabs relative to Americans (assessed by asking raters to complete feeling thermometers toward both groups). When assessing visual dehumanization in these studies, then, we were able to control not just for whether generators represent Arabs less favorably than Americans (e.g., Ratner et al., 2014), but also for the possibility that *raters' own*

[7] The standardized relation between these two rating measures spanned from β = .46–.62, depending on the rating study (all *p*s < .001). Although we did not pre-register our intention to collapse these measures into a single index of dehumanization, we do so here for the sake of efficiency; we arrive at the same conclusions if we treat these measures separately.

anti-Arab biases inflate the extent to which they rate representations of Arabs as more dehumanizing than representations of Americans.[8]

Do Americans, on Average, Mentally Represent Arabs in Dehumanizing Ways?

We first examined whether American generators, on average, mentally represent Arabs with blatantly dehumanizing features. In addition, we wanted to ensure that any mentally represented dehumanization of Arabs would not be reducible to a) a tendency for generators to mentally represent the outgroup less favorably than the ingroup, or to b) a tendency for raters to project their own dehumanizing views of Arabs onto the images they were rating. Determining the average degree to which Americans' mental representations of Arabs are dehumanized is useful insofar as it identifies a baseline for later examining whether—and if so, how much—the magnitude of mentally represented dehumanization varies across levels of generators' explicitly held beliefs (e.g., their explicit dehumanization of Arabs, or their levels of political conservatism).

Results from Group Image Rating Studies

We extracted overall group composite images of "Arabs" and "Americans" from each of the MTurk and Qualtrics Panels samples for raters to evaluate. These images reflect generators' averaged mental representations of Arabs and Americans, respectively (see Figure 2). All four of these representations were then rated, in a single rating study, by $N = 105$ MTurkers in a randomized order. We predicted (a) that representations of Arabs would be rated as more dehumanizing than representations of Americans, and (b) that this effect would hold when controlling for generators' tendency to represent Arabs more negatively than Americans, as well as for raters' own anti-Arab biases (that is, their prejudice against and dehumanization of Arabs).

[8] Of note, we found evidence that raters' actually *do* rate Arab images as more dehumanized when they themselves score higher in either explicit dehumanization of Arabs or in prejudice against Arabs (see supplementals).

To examine these predictions, we created a multilevel model in which we regressed ratings of how dehumanizing the representations appeared onto within-person contrasts that corresponded to our 2 (target: Arab, American) × 2 (image source: MTurk sample, Qualtrics Panels sample) factorial design. This model included random intercepts for each rater, which adjusted for the fact that the full-factorial design was nested within person. This model is conceptually analogous to running a 2 × 2 repeated measures ANOVA. Under this analytic strategy, we had 80% power to detect main effects as small as $\beta = 0.20$, and interactions as small as $\beta = 0.26$. These and all subsequent power estimates were derived by running Monte Carlo simulations on our models (which we did using the "simr" package: Green & MacLeod, 2016; see also Bolger, Stadler & Laurenceau, 2012).

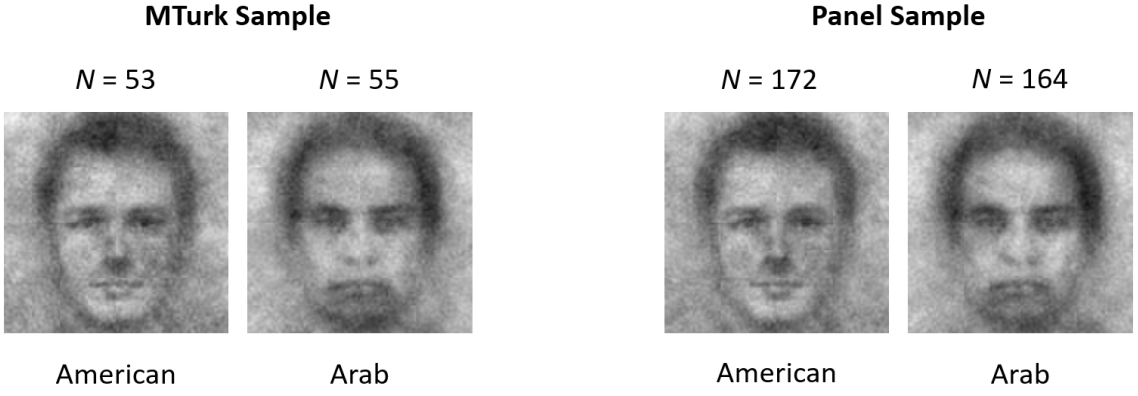

Figure 2. Generators' overall group composite images as a function of target condition ("Americans" vs. "Arabs") and generator sample (MTurk sample, Qualtrics Panels sample). These are the four images that raters saw in the study described in Row 1 of Table S1.

Confirming hypotheses, this analysis yielded a main effect of target, such that representations of Arabs were rated as substantially more dehumanizing than representations of Americans: $\beta = 1.24$, 95% CI[1.14, 1.34], $F(1, 313) = 543.22$, $p < .001$.[9] Of note, the Arab representations ($M = 28.97$, $SE = 1.61$) were also rated as looking less favorable (i.e., warm) than

[9] All multilevel models were constructed using the "lme4" package in R (Bates, Mächler, Bolker, & Walker, 2015). Degrees of freedom were estimated from the "lmerTest" package (Kuznetsova, Brockhoff, & Christensen, 2016), which uses the Satterthwaite approximation. Fluctuations in degrees of freedom are attributable to approximation variability.

representations of Americans (M = 68.34, SE = 1.61): M_{diff} = –39.47, β = –1.40, 95% CI [–1.51, –1.28], F(1, 307) = 535.55, p < .001. Importantly, however, dehumanization of Arabs held when controlling for ratings of how unfavorable the mental representations of Arabs (vs. Americans) appeared, as well as when controlling for raters' own anti-Arab biases (i.e., their prejudice against and blatant dehumanization of Arabs), though it did attenuate: β = 0.46, 95% CI[0.33, 0.59], F(1, 362) = 46.47, p < .001. This suggests that the dehumanization documented here is attributable to the views of *generators'* rather than to the views of raters. Figure 3 presents the total dehumanization effect on the left (with no covariates added to the model); on the right, this same effect is presented while covarying out how favorable Arabs (vs. Americans) were mentally represented as well as while controlling for raters' anti-Arab biases.

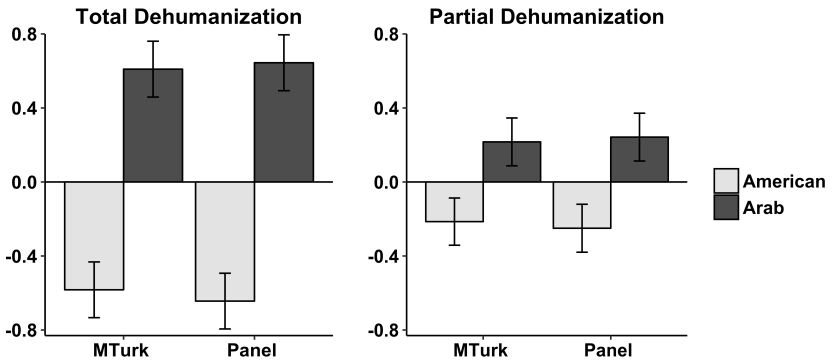

Figure 3. Z-standardized ratings of how dehumanizing mental representations appear, on average, as a function of who is being mentally represented (Arabs, Americans) and whose data were used to generate the representations (MTurk sample, Panels sample). The left-hand side of the figure depicts dehumanization ratings with no covariates added to the model; the right-hand side of the figure represents dehumanization ratings with covariates added. Error bars represent 95% confidence intervals.

Interestingly, this 2 (target: Arab, American) × 2 (sample: MTurk, Qualtrics Panels) analysis did not yield a main effect of sample (i.e., MTurk vs. Panel: β = –0.01, p = .80) or a target × sample interaction (interaction β = 0.10, p = .37). This suggests that the MTurk and Qualtrics Panels samples generated mental representations that were not distinguishable from the perspective of the people who rated them (at least not to a degree that we could reliably detect).

Results from Individual Image Rating Studies

We next analyzed ratings of generators' individual-level composite images. These individual composite images reflect the way that *individual generators* mentally represented Arabs vs. Americans. To analyze ratings of these representations, we constructed two multilevel models—one for each rating study (described in rows 6 and 7 of Table S1, respectively). Whereas generators were randomly assigned to call to mind either Arabs *or* Americans during the reverse-correlation task, raters evaluated even numbers mental representations from *both* of these conditions. In the multilevel models we constructed, ratings of how dehumanized the mental representations appeared were regressed onto a within-person contrast representing whether each representation was of an 'American' or of an 'Arab' (American = –½, Arab = ½). Both models included random intercepts for each rater, which adjusted for the fact that this factor was nested within person. These models also included random intercepts for each composite image, which adjusted for the fact that each participant saw a random subset of potential stimuli (i.e., a random subset of all the individual representations in the stimulus pool: Judd, Westfall, & Kenny, 2012). Because these models co-varied out the random effects of both raters and stimuli, they were very highly powered by conventional standards. Indeed, this study design and modeling strategy gave us 80% power to detect main effects as small as $\beta = 0.09$ for MTurk image rating study (which had $N = 397$ raters), and as small as $\beta = 0.038$ for the Qualtrics Panels image rating study (which had $N = 820$ raters). To put this into context, these effects are small enough to account for as little as 8.9 and 2.3 percent of the variance in rated dehumanization, respectively.[10]

[10] Variance-explained metrics (for these and all subsequent multilevel models) were computed by generating R^2 values for the referenced betas (for more on this technique, see Edwards, Muller, Wolfinger, Qaqish, & Schabenberger, 2008; see also Page-Gould, Sharples, & Song, 2019).

Results from the individual-level image rating studies corroborate those of the group image rating study described above. Specifically, we found that on average, individual representations of Arabs were rated as more dehumanizing than individual representations of Americans (MTurk sample: $\beta = 0.38$, 95% CI[0.24, 0.51], $F(1, 105) = 30.27$, $p < .001$; Qualtrics Panels sample: $\beta = 0.32$, 95% CI[0.26, 0.38], $F(1, 326) = 119.03$, $p < .001$). Moreover, this effect held when we controlled for the fact that Arabs were represented more negatively than Americans (in the case of the MTurk Sample: $\beta = 0.16$, 95% CI[0.11, 0.22], $F(1, 95) = 30.33$, $p < .001$) as well as when we controlled for represented negativity *and* raters' anti-Arab prejudice and dehumanization (in the case of the Qualtrics Panels sample: $\beta = 0.11$, 95% CI[0.09, 0.14], $F(1, 312) = 65.89$, $p < .001$). These analyses suggest that dehumanization is detectable in the mental representations of individuals as well as in the mental representations of groups of people, and they likewise suggest that dehumanization in the mind's eye is not reducible to 'mere' prejudice (on the part of the generators) or anti-Arab biases on the part of image raters.

Discussion

Our results consistently reveal that a low-status target group often subject to explicit blatant dehumanization is also mentally represented in blatantly dehumanizing ways: (non-Arab) generators' mental representations of 'Arabs' were rated not only as looking more negative than their representations of 'Americans' (the control group), but *specifically* as looking more dehumanizing. Importantly, we confirmed this research finding using ratings of composite images—proxies for mental representations (Dotsch & Todorov, 2012)—that were aggregated across groups of generators as well as at the level of individual generators. Moreover, because controlled for raters' own anti-Arab biases (i.e., their prejudice and dehumanization), we can be

more confident that any dehumanization that we observed is actually attributable to generators' representations rather than to biases that raters might impute onto those representations.

Our results are also noteworthy in providing new evidence for the stability and reliability of the composite images generated using the reverse-correlation technique. Our data revealed that there was no statistically distinguishable difference in the ratings of the group images obtained from the MTurk vs. Qualtrics Panels samples of generators—this despite the fact that the Qualtrics Panels sample had roughly three times as many generators than the MTurk sample and differed demographically. This observation suggests that a) group composite images using the reverse-correlation procedure can be quite reliable, and that b) $n = 50$ generators per image may suffice for obtaining a reasonably stable estimate of a group-aggregated mental representations.

Dehumanization in the Representations of "Low" vs. "High" Explicit Dehumanizers

Results thus far highlight the utility of using the reverse-correlation technique for capturing blatant dehumanization in a more indirect, arguably more implicit manner. Indeed, results confirm that Arabs are mentally represented in blatantly dehumanizing ways. Next, we sought to examine how our unobtrusive measure of dehumanization relates to more explicit (i.e., self-reported) forms of blatant dehumanization: Do individuals higher on explicit blatant dehumanization of Arabs hold more blatantly dehumanizing mental representations of Arabs, too? And if so, to what extent? Are blatantly dehumanizing representations of Arabs restricted to those who explicitly dehumanize them, or do they extend even to those who report low explicit dehumanization? We again assessed these questions using ratings of both group- and individual-level mental representations. As noted previously, there are a variety of advantages that come

with each of group-level moderation analyses and individual-level moderation analyses. We therefore report both to test our hypotheses as rigorously as possible.

Results from Group Image Rating Studies

To examine the question of whether generators' explicit blatant dehumanization moderates how dehumanizing their mental representations of Arabs appear, we conducted two rating studies (see Table S1, rows 2-3). In these studies, a total of $N = 395$ raters viewed four group-aggregated composite images in a 2 (target: Arab, American) × 2 (generator explicit dehumanization: low, high) × 2 (dehumanization measure type: trait, ascent) design with repeated measures on the first two factors. To create composite images—reflecting mental representations—from those "low" and "high" in explicit blatant dehumanization, we aggregated across generators who were below vs. above the median on self-reported dehumanization (for the MTurk sample), or across generators who were in the bottom vs. top quartile on self-reported dehumanization (for the Qualtrics Panels sample).[11] The third between-subjects factor—measure type—refers to whether raters saw mental representations that were aggregated by generators' dehumanization levels on the *trait* measure of explicit blatant dehumanization, or on the *ascent* measure of explicit blatant dehumanization. Because this third factor is not theoretically interesting and does not influence our conclusions, we do not discuss it further (but see the online supplement for a full discussion). Of note, generators in the "high" dehumanization groups self-reported substantially greater dehumanization of Arabs (vs. Americans) than did generators in the "low" dehumanization groups (all $\beta s \geq 1.51$, all $ps < .001$). For an illustration

[11] Each aggregated composite image therefore included the data of roughly $n = 25$ generators per image in the case of the MTurk Study, and roughly $n = 42$ generators per image in the case of the Panel Study. Readers can refer to the online supplement to see group composite images from this and all subsequent rating studies. We pre-registered our intention to split images at quartiles rather than medians in the Qualtrics Panels sample. This decision was motivated by a desire to obtain stable composite images that were as extreme as possible on explicit dehumanization.

of what group-level mental representations look like as a function of generators' levels of ascent- and trait-dehumanization of Arabs, respectively, see Figure 4.

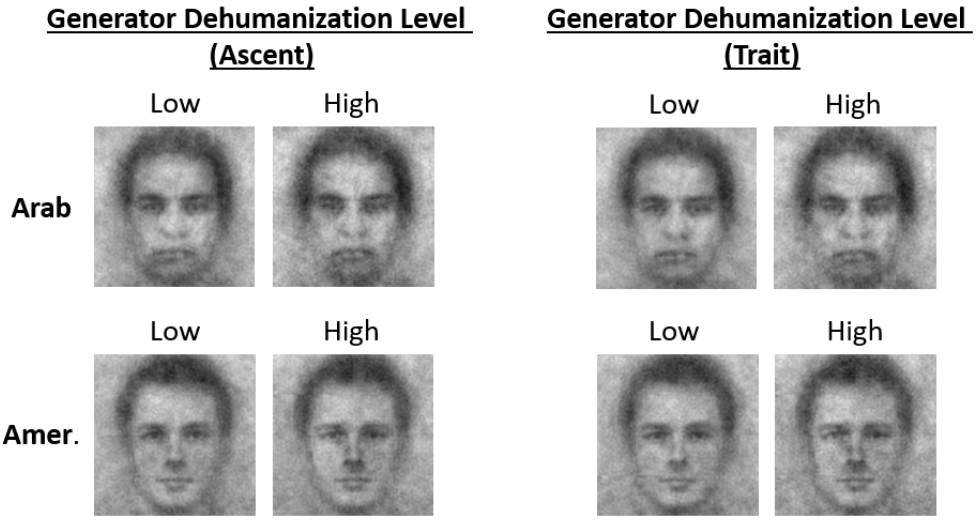

Figure 4. Group composite images broken down by generators' explicit dehumanization scores on each of two measures: ascent dehumanization and trait dehumanization. These images were generated by the Qualtrics Panels sample, and they were rated by participants in the rating study described in Row 3 of Table S1.

To analyze ratings from these studies, we constructed two multilevel models, one for each rating study. In these models we regressed ratings of how dehumanizing mental representations appeared onto contrast codes representing the full 2 × 2 × 2 mixed factorial design, as well as onto the mean-centered covariates referenced above (i.e., ratings of how favorably Arabs and Americans were mentally represented; raters' own anti-Arab prejudice and dehumanization).[12] These models also included random intercepts for each rater, which adjust for the fact that ratings of representations were nested within person. This experimental design and modeling strategy gave us 80% power to detect main effects as small as $\beta = 0.15$, and interaction effects between image target (Arab, American) and generator explicit dehumanization level (low vs. high) as small as $\beta = 0.20$.

[12] Throughout the remainder of this manuscript, we continue to control for how favorably Arabs vs. Americans were represented and, where possible, for the extent to which raters themselves are prejudiced against and dehumanize Arabs (vs. Americans). All conclusions hold regardless of whether these covariates are included.

Subjecting ratings of how dehumanized the mental representations appeared to the analysis described above yielded, as before, a main effect of target. Mental representations of Arabs were rated as more dehumanizing than mental representations of Americans (MTurk sample: $\beta = 0.47$, 95% CI[0.38, 0.57], $F(1, 669) = 89.32$, $p < .001$; Qualtrics Panels sample: $\beta = 0.50$, 95% CI[0.41, 0.60], $F(1, 692) = 104.63$, $p < .001$). Notably, this effect was *not* moderated by whether generators were high vs. low in self-reported dehumanization in the MTurk Sample (interaction $\beta = 0.04$, $p = .55$). Thus, in the MTurk sample both generators high *and* low in self-reported dehumanization had mental representations that were highly dehumanizing ("high" dehumanizers: $\beta = 0.49$, 95% CI[0.37, 0.61], $F(1, 641) = 66.44$, $p < .001$; "low" dehumanizers: $\beta = 0.45$, 95% CI[0.34, 0.57], $F(1, 639) = 57.64$, $p < .001$) and similarly so.

In the Qualtrics Panels sample, we found that generators' self-reported dehumanization levels (i.e., "low" vs. "high") significantly but weakly moderated the extent to which mental representations of Arabs were rated as dehumanizing: interaction $\beta = 0.13$, $F(1, 588) = 3.99$, $p = .046$, $R^2 = .01$. In this sample, representations of Arabs (vs. Americans) were rated as dehumanized to a greater degree when generated by "high" dehumanizers ($\beta = 0.57$, 95% CI[0.45, 0.69], $F(1, 590) = 88.42$, $p < .001$) as compared with when generated by "low" dehumanizers ($\beta = 0.43$, 95% CI[0.32, 0.55], $F(1, 655) = 55.52$, $p < .001$). Thus, there was at least some evidence that the effect was stronger among generators who were "high" in explicit dehumanization, but importantly, both "low" and "high" explicit dehumanizers mentally represented Arabs as substantially more dehumanized than Americans.[13]

[13] Of note, we did not pre-register the prediction that we would find significantly stronger dehumanization in the mental representations of "high" vs. "low" explicit dehumanizers. Because we found no evidence of this pattern in the MTurk sample, we expected the same would be true of the Qualtrics Panels sample. At the same time, this finding is broadly consistent with previous literature, which suggests that self-reported constructs can moderate the nature of mental representations captured using the reverse-correlation task (e.g., those higher on hostile sexism have more masculine-looking mental representations of feminists; Gundersen & Kunst, 2018).

Results from Individual Image Rating Studies

We next examined whether these conclusions held among ratings of individual-level composite images (that is, images that reflect individual generators' mental representations). To analyze ratings of individual representations, we constructed two multilevel models, one for each rating study (see Table S1, rows 6 and 7). In these models we regressed ratings of how dehumanizing the mental representations appeared onto (a) within-person contrasts representing whether representations were of Arabs vs. Americans; (b) generators' levels of explicit dehumanization (z-standardized at the level of generators); and (c) a contrast representing the cross-level interaction between these two variables. In addition, these models controlled for ratings of how favorable the mental representations appeared (in the case of the MTurk image rating study), as well as for raters' anti-Arab biases (in the case of the Qualtrics Panels image rating study). As before, these models included random intercepts for each rater as well as random intercepts for each mental representation (i.e., stimulus image). Again, because these models controlled for the random effects of raters *and* stimuli, they were very highly powered by conventional standards. These models had 80% power to detect interactions between target (Arab vs. American) and generators' dehumanization level as small as $\beta = 0.10$, in the case of the MTurk sample rating study (which had $N = 397$ raters), and as small as $\beta = 0.04$, in the case of the Qualtrics Panels sample rating study (which had $N = 820$ raters). For reference, this means we had 80% power to detect interactions that account for as little as 9.2 and 2.5 percent of the variance in rated dehumanization, respectively.

In the MTurk sample, we again found no evidence of significant moderation: the extent to which individual mental representations of Arabs were rated as more dehumanizing than those of Americans did not significantly vary as a function of generators' explicit dehumanization

levels (interaction β = –0.02, p = .63). The same was true of the Qualtrics Panels sample. In that sample, how dehumanizing Arab (vs. American) representations were rated also did not vary as a function of generators' explicit dehumanization levels: interaction β = 0.03, $F(1, 322) = 3.42$, p = .066, R^2 = .01. However, it is worth noting that although non-significant, this latter interaction effect did trend in the same direction as what we found in the group-rating analyses, above. Specifically, mental representations of Arabs (vs. Americans) were rated—directionally, albeit not significantly—as more dehumanizing when generated by those who were high (+1 SD) in explicit dehumanization (β = 0.14, 95% CI[0.10, 0.18], $F(1, 322) = 50.57$, $p < .001$) than when generated by those who were low (–1 SD) in explicit dehumanization (β = 0.09, 95% CI[0.05, 0.13], $F(1, 313) = 20.29$, $p < .001$). Still, generators' explicit dehumanization levels did not moderate the magnitude of this effect to a significant degree—suggesting that low vs. high explicit dehumanizers have mental representations of Arabs that are both heavily dehumanized, and more similar than they are different.

Discussion

Our analyses suggest that the tendency to mentally represent Arabs in dehumanizing ways may indeed be slightly greater among people who self-report greater levels of explicit dehumanization of Arabs. On the one hand, this helps provide some evidence of convergent validity, insofar as it suggests a link between explicit dehumanization and the tendency to mentally represent a group of people in blatantly dehumanizing ways. On the other hand, our analyses suggest that these two forces are by no means redundant; even those who explicitly dehumanize Arabs *the least* harbor mental representations of Arabs that are significantly more dehumanizing than their mental representations of Americans. In fact, the representations of those who strongly dehumanize Arabs on explicit measures are remarkably similar to the

representations of those who weakly dehumanize Arabs on explicit measures. In more cases than not, differences between the mental representations of low vs. high explicit dehumanizers were non-significant, even under the scrutiny of highly powered statistical testing. This suggests that relying on explicit measures of blatant dehumanization alone may lead researchers to underestimate the degree to which a population dehumanizes a target group.

Dehumanization in the Mental Representations of Liberals vs. Conservatives

We next turned to consider the question of whether dehumanization in one's mental representations varies across the political spectrum. There is evidence suggesting that, at the explicit level, political liberals report less blatantly dehumanizing views of low-status groups than do political conservatives (e.g., Costello & Hodson, 2007; Esses, Veenvliet, Hodson, & Mihic, 2008; Kteily et al., 2015; Maoz & McCauley, 2008). But would any clear distinction remain when looking more unobtrusively at individuals' mental representations? Examining the extent to which dehumanizing mental representations vary across the political spectrum matters, even beyond examining moderation as a function of individuals' own explicitly dehumanizing attitudes—if people incorrectly consider blatant dehumanization of low-status groups to be largely the province of those on the political right, it would substantially (and artificially) limit our sense of the scope of the problem.

On the one hand, there is reason to believe that political liberals would have mental representations of Arabs that are less dehumanized than those of political conservatives, matching their more humanizing self-reports. For example, liberals tend to express less *prejudice* against low-status groups than conservatives do, even on implicit measures. On the implicit association test, for instance, people who are extremely liberal have reliably lower *D*-scores (when it comes to evaluative IATs assessing biases against low-status groups) than people who

are extremely conservative (though notably, this gap is smaller than the difference at the explicit level; Nosek et al., 2009). On the other hand, our findings in the previous section suggest that even individuals who self-report lower levels of explicit blatant dehumanization can have overtly dehumanizing mental representations that are highly comparable to those among individuals who self-report high levels. If explicit beliefs do little to moderate one's mental representations, then as a group, liberals may harbor representations of Arabs that are *as dehumanizing* as those of conservatives.

Results from Group Image Rating Studies

To examine the question of whether generators' political orientation moderates how dehumanizing their mental representations of Arabs appear, we conducted two rating studies (described in rows 4 and 5 of Table S1). In these studies, a total of $N = 584$ raters viewed four group composite images (that is, group representations) in a 2 (target: Arab, American) × 2 (generator ideology: left vs. right wing) × 3 (ideology measure: one-item liberalism-conservatism, SDO, RWA) design with repeated measures on the first two factors. To create representations for "left wing" versus "right wing" generators, we aggregated across generators who were below vs. above the median on our political ideology measures (in the case of the MTurk sample) or in the bottom vs. top quartile (in the case of the Qualtrics Panels sample).[14] The third between-subjects factor—ideology measure—refers to whether raters saw mental representations that were grouped by generators' political orientation levels on the one-item measure of political ideology (which spanned from *0 = extremely liberal* to *10 = extremely conservative*), on the SDO measure of political ideology, or on the RWA measure of political

[14] Here, as before, we pre-registered our intention to split composite images by quartiles rather than medians in the Qualtrics Panels sample. Again, this decision was motivated by a desire to obtain stable composite images that were as extreme as possible on the dimensions of interest (in this case, generators' political leanings), in order to conduct a more conservative test of our hypotheses.

ideology. Of note, generators in the right-wing groups self-reported substantially more right-leaning attitudes than those in the left-wing groups across all of our measures of political ideology (all βs ≥ 1.67, all *p*s < .001). In addition, and consistent with past research, generators in the right-wing groups reported substantially greater explicit dehumanization of Arabs (vs. Americans) than those in the left-wing groups (all βs ≥ 0.44, all *p*s < .001). Indeed, generators who were grouped as left-wing explicitly dehumanized Arabs only to a very small degree, if at all (see supplementals for specifics). For an illustration of what group-level mental representations look like as a function of whether they were created by generators who were politically left- vs. right-wing, respectively, see Figure 5.

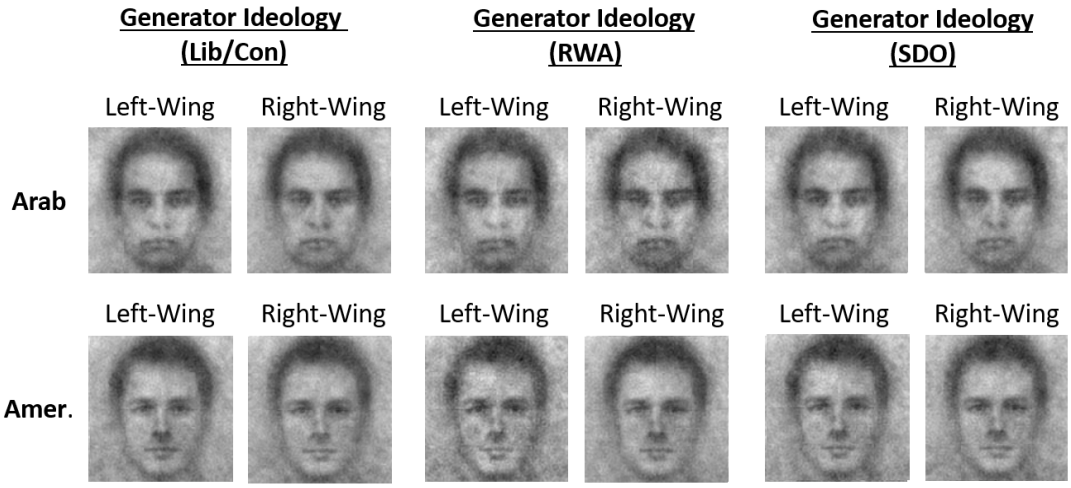

Figure 5. Group composite images broken down by generators' political ideologies on each of three measures: 1-item liberalism/conservatism, RWA, and SDO. These images were generated by the Qualtrics Panels sample, and they were rated by participants in the rating study described in Row 5 of Table S1.

To analyze ratings from these studies, we constructed two multilevel models, one for each rating study (i.e., those in rows 4 and 5 of Table S1). In these models we regressed ratings of how dehumanizing the group representations appeared onto contrast codes representing the full 2 × 2 × 3 mixed factorial design, as well as onto the mean-centered covariates described

previously (how favorable the people in the mental representations appeared, in the case of the MTurk sample rating study; favorability as well as raters' anti-Arab biases, in the case of the Panels sample rating study). These models included random intercepts for each rater, which adjusted for the fact that ratings of mental representations were nested within person. This analytic strategy gave the studies 80% power to detect main effects as small as $\beta = 0.12$, and 80% power to detect interactions between target (Arab vs. American) and generator ideology (left-wing vs. right-wing) as small as $\beta = 0.17$.

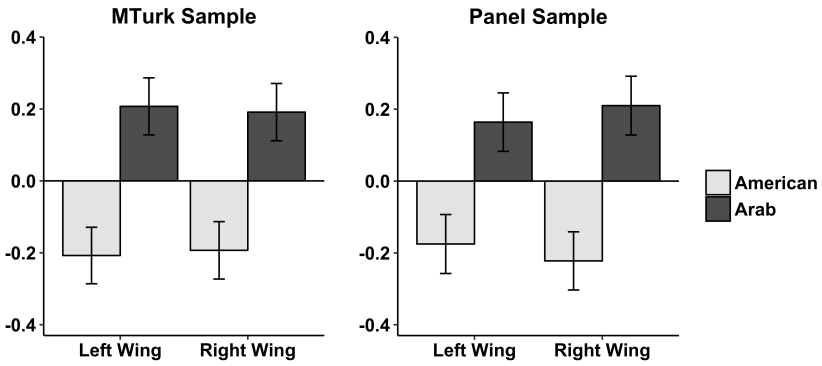

Figure 6. Z-standardized ratings of how dehumanizing mental representations appear, on average, as a function of who is being mentally represented (Arabs, Americans) and whether the people harboring the representations are left-wing vs. right-wing. Data from the MTurk sample are depicted on the left; data from the Panels sample are depicted on the right. Error bars represent 95% confidence intervals.

Subjecting ratings of mental representations to this analysis replicated the main effect of the target that we observed previously. Mental representations of Arabs were rated as more dehumanizing than mental representations of Americans (MTurk sample: $\beta = 0.40$, 95% CI[0.32, 0.48], $F(1, 1011) = 105.25$, $p < .001$; Qualtrics sample: $\beta = 0.39$, 95% CI[0.31, 0.47], $F(1, 980) = 90.08$, $p < .001$). In the MTurk sample, this effect was not moderated by whether generators were left-wing vs. right-wing (interaction $\beta = –0.03$, $p = .57$). In the Qualtrics Panels sample, the results looked much the same. That is, the tendency for representations of Arabs to be rated as more dehumanizing than representations of Americans did not depend on the political ideology

of those who generated them (interaction $\beta = 0.09$, $p = .11$; see Figure 6). Moreover, these effects were not qualified by which ideology measure we used when grouping generators as left- vs. right-wing (all three-way interaction $ps \geq .09$; but see supplemental materials for more information on this non-significant three-way interaction).

Results from Individual Image Rating Studies

To analyze ratings of individual composite images (rows 6-7 of Table S1), we constructed multilevel models in which we regressed ratings of how dehumanizing individual representations appeared onto (a) within-person contrasts representing whether representations were of Arabs vs. Americans; (b) generators' political ideology levels (either generator RWA, SDO, or liberalism-conservatism level, depending on the model; always z-standardized at the level of generators); and (c) a contrast representing the cross-level interaction between these two factors. In addition, these models controlled for ratings of how favorably individual mental representations appeared (in the case of the MTurk image rating study), as well as for raters' anti-Arab prejudice and dehumanization (in the case of the Qualtrics Panels image rating study). As before, these models included random intercepts for each rater as well as random intercepts for each mental representation (that is, each individual composite image). These models had 80% power to detect interactions between target (Arab vs. American) and generators' ideology level as small as $\beta = 0.10$ in the case of the MTurk sample image rating study (which had $N = 397$ raters), and as small as $\beta = 0.04$ in the case of the Qualtrics Panels image rating study (which had $N = 820$ raters). Again, this means we had 80% power to detect interactions that account for as little as 9.2 and 2.5 percent of the variance in rated dehumanization, respectively.

When looking at ratings of individual representations from the MTurk sample, the tendency for representations of Arabs to be rated as more dehumanizing than representations of

Americans was not moderated by generators' levels of RWA (interaction β = 0.02, p = .59) or SDO (interaction β = 0.02, p = .54). However, it *was* significantly (albeit modestly) moderated by where generators placed themselves on the 1-item liberalism-conservatism measure (interaction β = 0.07, 95% CI[0.01, 0.13], $F(1, 93) = 4.47$, p = .037, R^2 = .05). Here, generators who were a standard deviation more conservative than the mean mentally represented Arabs (vs. Americans) in ways that were more dehumanizing (β = 0.23, 95% CI[0.14, 0.31], $F(1, 90) = 25.42$, $p < .001$) than generators who were a standard deviation more liberal than the mean (β = 0.09, 95% CI[–0.00, 0.18], $F(1, 91) = 3.65$, p = .059). However, this finding did not replicate in the larger (and more representative) Qualtrics Panels study. There, generators' political ideology did *not* moderate the extent to which their mental representations of Arabs were rated as more dehumanizing than their mental representations of Americans—and this was true regardless of how political ideology was operationalized (see Figure 7). That is, in the Qualtrics Panels study we found *no evidence* that the tendency to mentally represent Arabs (vs. Americans) in dehumanizing ways was moderated by generators' levels of RWA (interaction β < 0.01, p = .76), SDO (interaction β < 0.01, p = .89), or self-placement on the 1-item measure of liberalism vs. conservatism (interaction β < 0.01, p = .72). Moreover, follow-up equivalence tests suggested that all three of these interaction betas were *significantly* closer to zero than they were to the smallest interaction betas we had 80% power to detect (all equivalence test ps $\leq$ .007).[15] This suggests that in the Qualtrics Panels study, generators' political ideology had virtually zero impact on how much they mentally represented Arab (vs. American) targets in dehumanizing ways.

[15] Equivalence tests were conducted by performing two one-sided t-tests of our observed interaction effects against interaction βs of 0.04 and –0.04 (the least non-zero interactions we were well-powered to detect; see Lakens, 2017).

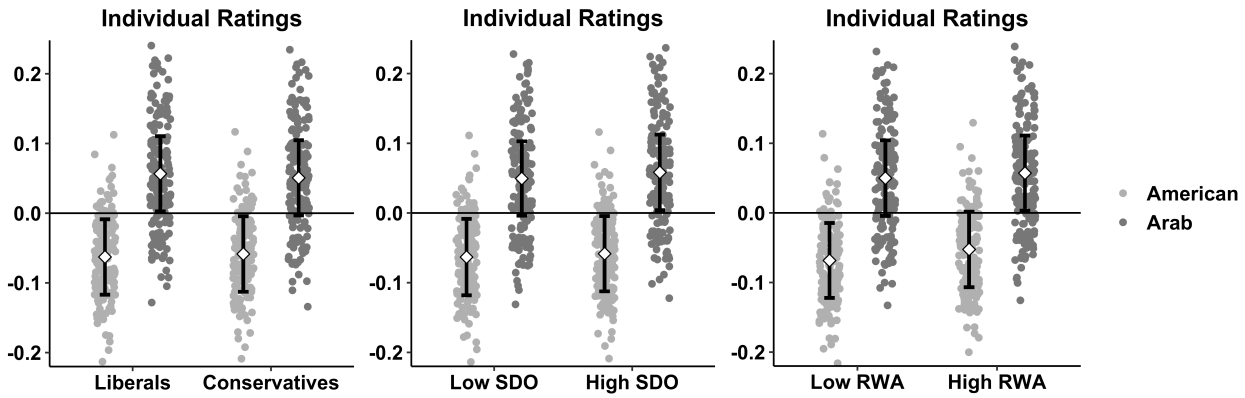

Figure 7. Z-standardized ratings of how dehumanizing representations appear as a function of whether the people who generated them were a standard deviation more liberal or more conservative than the sample mean on each of three measures of political ideology (one-item liberalism-conservatism; SDO; and RWA). Each dot represents the averaged dehumanization rating of an individual representation from the Qualtrics Panels sample. Diamonds represent marginal condition means, which are encompassed by 95% confidence intervals.

Discussion

In sum, we found little evidence that generators' political ideology influenced how dehumanizing their mental representations of Arabs (vs. Americans) appeared. Instead, our results suggested that, despite liberals standing apart from conservatives in eschewing explicit blatant dehumanization of Arabs, both liberals and conservatives harbor mental representations of Arabs that are similarly blatantly dehumanized (as assessed using the unobtrusive reverse correlation method). Importantly, in the present studies, the absence of evidence for moderation cannot readily be explained by low statistical power to observe interaction effects. Indeed, our most highly powered tests of moderation revealed that our observed interaction betas were significantly closer to zero than they were to interaction βs as small as 0.04 and –0.04, respectively—that is, interaction betas that are small enough to account for just 2.5% of the variance in dehumanization ratings.

These findings are noteworthy for at least two related reasons. First, they extend existing work on liberals and conservatives' self-reported vs. indirectly-assessed perceptions of low-

status groups, which tends to focus on *attitudes*; here, we consider ideological differences using more versus less obtrusive measurement in intergroup perceptions that are specifically *dehumanizing* and that are not reducible to mere valence (see also Bruneau et al., 2018). Second, our findings differ from previous work which suggests that, despite a smaller gap at the implicit level, liberals evince less bias than conservatives on both self-report and more indirect measures (e.g., Nosek et al., 2009). Here, in contrast, we find that despite clear ideological differences on explicit blatant dehumanization, the tendency to mentally represent Arabs in blatantly dehumanizing ways is *just as prevalent* in the minds of liberals as it is in the minds of conservatives. Importantly, this suggests that blatant dehumanization of low-status groups may—at least when it comes to dehumanizing mental representations—apply much more broadly than previously assumed.

Do Generators Have Insight into How Dehumanizing Their Mental Representations Are?

The above analyses suggest that blatantly dehumanizing representations of Arabs can be just as prevalent among individuals exhibiting low levels of explicit dehumanization (e.g., liberals) as among individuals exhibiting high levels of explicit dehumanization (e.g., conservatives). This raises an important question: Are individuals who self-report low levels of blatant dehumanization *aware* that they harbor such dehumanizing mental representations? In a final set of exploratory analyses, we begin to consider whether generators have insight into the degree of dehumanization in their minds' eyes. On the one hand, generators who exhibit low levels of explicit dehumanization may be unaware of just how dehumanizing their own mental representations of Arabs are and may be horrified to learn of it. On the other hand, it is conceivable that generators who exhibit low levels of explicit dehumanization have at least some awareness of how dehumanized the representations they harbor are—dehumanization that they

effortfully control or perhaps intentionally underreport because of social desirability concerns (Gawronski & De Houwer, 2014). This matters—if people who explicitly reject dehumanization are unaware that they harbor highly dehumanizing mental representations, then making them aware might motivate them to grapple with that dehumanization (e.g., see Monteith & Mark, 2005, for analogous reasoning). If, on the other hand, they *are* fully aware, more work would be needed to determine whether their lack of self-reported blatant dehumanization reflects 'honest' efforts to control negative explicit attitudes or a mere self-presentation strategy to conceal dehumanization that is internally endorsed (two possibilities that would have differing implications for the perniciousness of blatant dehumanization in society).

Exploratory Measures of Generator Insight

As mentioned previously, generators in the Qualtrics Panels sample completed exploratory measures that generators in the MTurk sample did not (see OSF for full details). Among these measures were two indices that allow us to (tentatively) explore the extent to which generators might have insight into how dehumanizing their mental representations are. We call these measures the *image-selection* and *percentile* measures of insight, respectively.

Image-selection measure of insight. To determine whether generators are aware of the dehumanization in their minds' eyes, we included an exploratory measure in the Qualtrics Panels study that we call the image-selection measure. This measure was accompanied by the following instructions:

> We are going to create an image morph of all the faces you chose during the task from earlier (a morph of all the faces you thought looked Arab [American]). When we do, which of the following images do you think your morph will most closely resemble? Please try and be as accurate as possible.

To incentivize accuracy (and, thereby, honest responding), generators were further told that the most accurate respondent would be given a $25.00 gift card to Amazon.com. Below these instructions, generators each saw five images (depicted in Figure 8). These images were all actual individual-level composite images of "Arabs" or "Americans," respectively, that had been obtained from the MTurk sample of generators. They spanned from the least dehumanizing mental representation in that sample (on the left) to the most dehumanizing mental representation in the sample (on the right), identified on the basis of the rating study described in row 6 of Table S1; the three middle images were, in order, mental representations that were at the 25th, 50th, and 75th percentile on rated dehumanization. We tested insight by examining whether individuals who predicted that they would have a less (vs. more) dehumanized mental representations on the image-selection measure *actually* had less dehumanized mental representations.

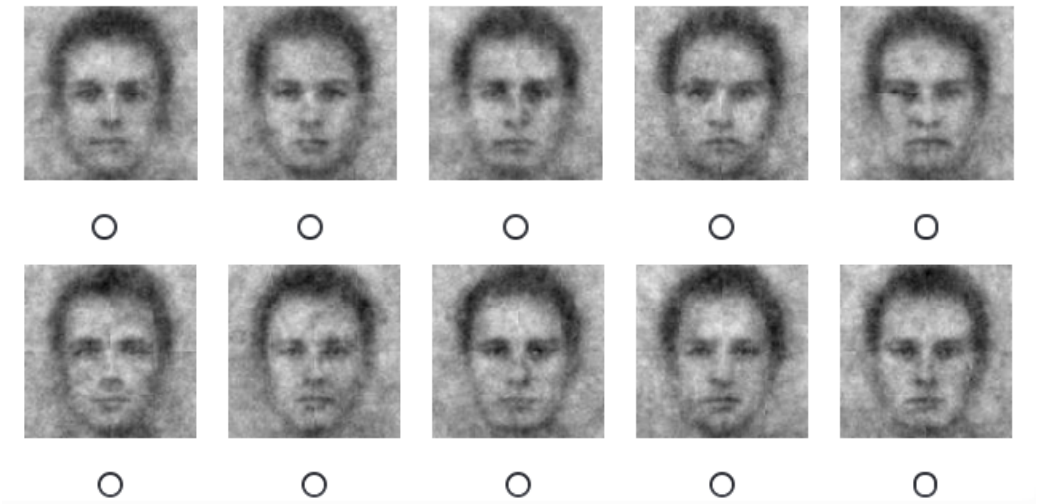

Figure 8. Arrays of images that were shown to generators in order to assess their anticipated levels of represented dehumanization. All images were individual-level images from the MTurk study. Those on the top are individual images of "Arabs"; those on the bottom are individual images of "Americans." Images span from least- (left) to most (right) dehumanized.

Percentile measure of insight. We also examined generators' levels of insight by including an exploratory measure that we call the percentile measure, which was always asked

directly after the image-selection measure. This measure was accompanied by the following instructions:

> What percent of adults living in the U.S. do you think has a *more negative* mental image of Arabs [Americans] than you do? Note that higher numbers mean you think *more* U.S. adults have more negative mental images than you.

Again, to incentivize accuracy, generators were told that the person who was most accurate at guessing their percentile (that is, their actual standing relative to other generators in the Panels sample) would receive a $25.00 gift card to Amazon.com. Generators provided their answers on a 100-point sliding scale from 0 to 100, with 100 corresponding to the belief that 100% of U.S. adults have a more negative mental image than generators. To turn this measure into a percentile projection, we reverse-scored generators' answers. For example, a generator who thought that only 20% of U.S. adults would have a more negative mental representation was considered to have placed themselves at the 80th percentile on mental representation negativity. We tested insight by considering whether those who anticipated being at a lower (vs. higher) percentile on mental representation negativity actually had mental representations that were rated as less dehumanized.

Exploratory Findings Related to Insight

If generators have introspective access into the degree of dehumanization in their own mental representations, then those who anticipate greater (vs. lower) levels of dehumanization on these exploratory measures should have generated representations of their own that were rated by others as more (vs. less) dehumanizing. In order to examine whether this was indeed the case (and whether insight differed among individuals among higher vs. lower explicit dehumanizers), we regressed ratings of how dehumanized individual representations appeared (collected from

the study described in Row 7 of Table S1) onto (a) generators' anticipated dehumanization levels (either on the image-selection measure or the percentile measure, depending on the model; always z-standardized at the level of the generators); (b) generators explicit beliefs (e.g., their self-reported levels of blatant dehumanization, or their self-reported levels of SDO; always z-standardized at the level of generators; and onto (c) interactions between these two variables. As in previous models, we included the same covariates that we have been using across all rating studies (that is, how favorably individual representations were rated, as well as raters' own anti-Arab prejudice and dehumanization levels). These models included random intercepts for each rater as well as random intercepts for each mental representation (that is, each individual composite image that was rated). According to Monte Carlo simulations, these models were powerful enough (that is > 80% powered) to detect standardized relationships between generators' anticipated and actual levels of dehumanization (in their mental representations) as small as $\beta = 0.03$.

When we subjected image ratings to the analysis described above, we found a weak association between anticipated and actual levels of dehumanization in the mind's eye. Specifically, generators who selected a more dehumanizing representation on the image-selection measure did indeed tend to harbor more dehumanizing mental representations, but only to a very small degree: $\beta = 0.02$, 95% CI[0.01, 0.04], $F(1, 322) = 9.97$, $p = .002$. Moreover, the magnitude of this weak association was not moderated by generators' explicit dehumanization levels ($\beta < 0.01$, $p = .98$). Thus, all generators, regardless of whether they were more or less likely to dehumanize at the explicit level, tended to have little accuracy at anticipating just how dehumanizing their own representations would appear—at least on the image-selection measure. When we examined moderation by generators' ideology instead of explicit dehumanization, we

again found little evidence that the (low) degree of insight depended on political liberalism vs. conservatism ($\beta < 0.01$, $p = .69$), RWA levels ($\beta = 0.01$, $p = .44$), or SDO levels ($\beta = 0.01$, $p = .09$).

Turning to the percentile measure, generators who anticipated being in a higher percentile on mental representation negativity tended to harbor mental representations of their own that were, if anything, *less* dehumanizing than those of generators who anticipated being in a lower percentile (though notably, this relationship was not significantly different from zero): $\beta = –0.01$, 95% CI[–0.03, 0.00], $F(1, 319) = 2.79$, $p = .10$. Again, this is to say that generators appeared to have little ability to anticipate how dehumanizing their own mental representations would be. As in the case of the analysis above, the magnitude of this (null) relationship was not moderated by generators' explicit dehumanization levels ($\beta < 0.01$, $p = .61$), RWA levels ($\beta = 0.01$, $p = .11$), or SDO levels ($\beta = 0.01$, $p = .50$). There was some evidence that this relationship was moderated by the 1-item measure of political liberalism vs. conservatism ($\beta = 0.02$, $p = .014$); however, the nature of this interaction was simply that the systematic lack of accuracy we observed overall was slightly more pronounced among liberals (–1 SD on the 1-item measure: $\beta = –0.03$, 95% CI[–0.05, –0.01], $F(1, 317) = 8.76$, $p = .003$) than among conservatives (+1SD on the 1-item measure: $\beta = 0.01$, 95% CI[–0.01, 0.03], $F(1, 320) = 0.41$, $p = .52$).

Across all analyses, then, we found that generators had a hard time anticipating how dehumanizing their mental representations would appear—even when they were monetarily incentivized to be as accurate as possible. Moreover, this lack of accuracy held across varying levels of explicit dehumanization, and it held as well across the political spectrum.[16]

[16] These analyses collapse across generators in both the American and Arab conditions; all conclusions hold if we focus our analyses only on generators who were in the Arab condition.

Another way to get at the question of insight is to investigate whether generators who anticipated being in a higher percentile were, in fact, likely to generate a mental representation that was in a higher percentile on rated dehumanization. In order to investigate whether this was indeed the case, we computed the actual percentile into which each generator's mental representation fell (on rated dehumanization; relative to other generators in the Panels sample), and we regressed this onto generators' *anticipated* percentile. Here, as before, we found limited evidence that generators have insight into the degree of dehumanization in their own mental representations. That is, generators' anticipated percentiles were unrelated to their actual percentiles: $\beta = 0.01$, 95% CI[–0.09, 0.12], $F(1, 334) = 0.06$, $p = .81$. This is to say that generators' guesses about how negative their mental representations would look tended to be unpredictive of how dehumanizing their mental representations ended up looking to naïve raters. Moreover, and as reported above, this pattern was not moderated by generators' explicit dehumanization levels ($\beta = 0.02$, $p = .68$), SDO levels ($\beta = 0.02$, $p = .67$), RWA levels ($\beta = 0.10$, $p = .06$), or political liberalism vs. conservatism ($\beta = 0.10$, $p = .08$) levels.[17] Thus, across these analyses, we again found that generators' accuracy levels were low, and that they tended to be low across varying levels of generators' explicit beliefs.[18]

Thus far, we have considered whether participants' expectations of the degree to which their mental representations are dehumanized match the levels of dehumanization that actually appear in their mental representations. In a final set of analyses, we looked instead at whether

[17] Decomposing these latter two interaction effects—which were not significant but were close the cutoff for statistical significance—does not change our overall conclusions. Instead, doing so reveals that both liberals' and conservatives' actual percentiles were uncorrelated with their anticipated percentiles, albeit to slightly differing degrees (e.g., –1 SD on RWA: $\beta = –0.07$, 95% CI[–0.21, 0.07], $F(1, 332) = 0.99$, $p = .32$; +1 SD on RWA: $\beta = 0.12$, 95% CI[–0.03, 0.28], $F(1, 332) = 2.42$, $p = .12$).

[18] Of note, these analyses do not imply that individuals have no idea what Arabs look like in their minds' eyes. That is, when completing the reverse correlation task, participants are presumably calling upon their mental representation of what Arabs look like, and it is because this mental representation is not random that the composite image that emerges looks Arab-like. Rather, our analyses more specifically suggest that participants have relatively little insight into the degree to which their mental representation of Arabs is specifically dehumanizing.

generators' explicit beliefs predicted these expectations in the first place— that is, how dehumanizing did individuals *expect* their mental representations of Arabs to be? Did generators who expressed less (vs. more) explicit dehumanization—or who were more left (vs. right)-leaning—*anticipate* having less (vs. more) dehumanizing representations of Arabs? On the image-selection measure, the answer appears to be 'yes.' That is, those who were lower in blatant dehumanization and who were more left-wing did generally tend to anticipate having less dehumanizing representations of Arabs ($\beta_{blatant} = 0.15$, $p = .031$; $\beta_{Lib\text{-}Con} = 0.08$, $p = .29$; $\beta_{SDO} = 0.25$, $p < .001$; $\beta_{RWA} = 0.14$, $p = .056$), and this despite our overall pattern of evidence (reported earlier) suggesting that their representations were typically just as dehumanizing. This pattern of findings accords with the possibility that more egalitarian generators may be unaware of the extent to which they harbor dehumanizing representations of Arabs. Of note, however, the results on the percentile measure offered additional perspective. On this measure, we found that those who were lower in explicit blatant dehumanization or who were more left-leaning predicted that they would have mental representations of Arabs that were no less negative than those predicted by those higher in explicit blatant dehumanization or more right-leaning ($\beta_{blatant} = –0.04$, $p = .55$; $\beta_{Lib\text{-}Con} = 0.06$, $p = .38$; $\beta_{SDO} = 0.08$, $p = .28$; $\beta_{RWA} = 0.01$, $p = .85$). Although tentative, these latter findings suggest that generators who report low levels of blatant dehumanization may be aware—at least to some degree—that their mental representations of Arabs are more dehumanizing than would be suggested by the humanity attributions they explicitly report.

Discussion

The analyses described in this section are exploratory (and therefore tentative), but they provide initial leverage on the question of whether people who self-report low levels of explicit dehumanization (e.g., liberals) are aware of the degree to which their mental representations of

Arabs are dehumanizing. At the broadest level, these analyses do not provide strong evidence that generators can accurately infer the levels of dehumanization in their own mental representations. Generators who anticipated harboring less dehumanizing mental representations did not consistently generate mental representations of their own that were, in line with their guesses, less dehumanizing. Finally, there was some—albeit mixed—evidence that those who explicitly rejected dehumanization of Arabs anticipated that they would be less likely to harbor dehumanizing mental representations of Arabs, even though most of the evidence we report in our paper suggests that they actually had mental representations that were about as dehumanizing as those higher on explicit dehumanization. In short, and despite the need to investigate this further in future research, the totality of the evidence reported in this section provides little support for the idea that those who reject explicit dehumanization are aware of the high levels of dehumanization that we observe in their mental representations.

General Discussion

The present paper validates the use of reverse-correlation image classification for indexing blatant dehumanization, and it suggests that this task can be used to capture dehumanization even in the minds of those who outwardly reject it. Moreover, this paper finds that dehumanization in one's mental representations is not reducible to (a) a general tendency for outgroup members to represented less favorably than ingroup members (Ratner et al., 2014), or to (b) a tendency for raters to project their own views (prejudices, dehumanization) onto the representations they rate. Further, this paper reports that dehumanization in the mind's eye is not redundant with explicit dehumanization. Although the two may be related to each other in theoretically reasonable ways, much variance in how dehumanizing one's mental representations are is left unexplained by where one falls on self-reported dehumanization measures. Finally, this

paper suggests that those who are politically liberal not only harbor more blatant dehumanization of an outgroup (here, Arabs) than they are willing to self-report, but that they may harbor *as much* blatant dehumanization—at least in terms of their mental representations—as those who are politically conservative. This pattern of findings is notable given that blatant dehumanization of low-status groups is often considered largely the domain of the political right (e.g., Jackson & Gaertner, 2010; Kteily & Bruneau, 2017b). Moreover, this finding stands out given that, as noted previously, previous research examining prejudice (e.g., anti-Black prejudice) consistently finds that liberals and conservatives tend to differ in their levels of outgroup bias, even on more indirect and implicit measures (Nosek et al., 2009).

We concluded with exploratory findings that relate to an important theoretical question—the question of whether people who self-report low levels of blatant dehumanization (e.g., liberals) are *aware* of how dehumanizing the mental representations they harbor are. Generally speaking, these analyses revealed that generators have a hard time anticipating how dehumanizing their own mental representations will appear, and that generators who self-report lower levels of blatant dehumanization tend to anticipate less dehumanization in their mental representations overall (though not always). These findings are aligned with the possibility that generators are unaware of how dehumanizing the images in their minds are, but these conclusions are tentative and are in need of more systematic confirmation (including a broader variety of measures of insight than we considered here). If it is the case, for example, that generators fail to realize just how dehumanizing their own mental representations can be, then informing egalitarian generators of the dehumanized mental representations in their minds could motivate them to grapple with and try and change the content of those representations (for example, by purposefully exposing themselves to more humanizing portrayals of Arabs). Toward

this end, future work should also seek to understand to what extent mental representations can be modified by top-down cognitive control efforts (see Bargh, 1994, for similar considerations in the automaticity literature). To the extent that we develop an understanding of these aspects of mental representations, we can perhaps find ways to change their content—and in turn, to reduce the prevalence and, perhaps, the consequences of intergroup dehumanization.

Indeed, a major implication of these findings is that blatant dehumanization may be more widespread than previously thought. If liberals and those who are "low" on explicit dehumanization nevertheless harbor clearly dehumanizing mental representations of Arabs, existing estimates of the dehumanization of marginalized groups (e.g., Kteily et al., 2015)—as striking as they are—may in fact underestimate the degree to which people consider members of these groups to be less than fully human. The perspective offered in this paper expands the scope of blatant dehumanization from perceivers explicitly reporting dehumanizing views about another group to perceivers harboring representations of another group (whether at explicit or more implicit levels) that others readily identify as dehumanizing.

Still, determining the implications of dehumanization in mental representations requires further empirical attention. It remains unclear to what extent and under what circumstances dehumanizing mental representations have downstream consequences. To the extent that mental representations revealed by reverse-correlation procedures operate like implicit measures of intergroup impressions, it stands to reason that, like implicit measures, variation in mental representations may be predictive of variation in intergroup behavior (for example, with more dehumanizing mental representations predicting greater hostility or aggression toward the relevant targets). That said, although there is good evidence for links between implicit attitudes and behavior, these links are not always as strong as researchers might expect (for a recent meta-

analysis, see Kurdi et al., 2019), and their predictive utility often depends on a variety of factors. For example, some work suggests that the strength of the links depends on whether behavioral criteria and implicit measures are specified in similar ways (Bodenhausen & Petsko, in press). Other work suggests that implicit measures may be more predictive of behavior when looking at aggregated levels of bias (e.g., at the level of counties or states) rather than when looking at individual-level variation (Payne, Vuletich, & Lunberg, 2017). Thus, even if mental representations do, like other implicit attitudes, predict downstream behavior beyond explicit measures, fully explicating when and why this is the case is likely to require more work. Of note, some emerging evidence does point to the consequentiality of mental representations in predicting intergroup behavior. For example, the tendency to mentally represent God as a White man—which itself has been captured using reverse-correlation procedures (Jackson, Hester, & Gray, 2018)—is associated with perceiving White individuals as more fit for leadership positions than Black individuals (Roberts et al., 2020), even when researchers control for participant's explicit levels of racism. Similarly, although outside the context of reverse correlation, Goff et al. (2014) found that beyond explicit anti-Black attitudes, the implicit mental association between African Americans and apes predicted overestimating the age of juvenile Black defendants and perceiving them as more culpable for their crimes. We await further research to address the important question of how and under what circumstances mental representations guide perceivers' intergroup behavior.

Future work should also directly investigate why mental representations of Arabs vs. Americans are relatively unmoderated by explicit beliefs, whereas representations of other groups, like 'feminists' or 'the poor', are (Gundersen & Kunst, 2018; Lei & Bodenhausen, 2017). One possibility is that political liberals actively conceal explicitly dehumanizing beliefs

they privately endorse about Arabs whereas they *actually* privately reject dehumanizing beliefs about groups like feminists or the poor. Another possibility is that liberal Americans' lack of explicit dehumanization of Arabs is genuine, but—because they lack direct contact with Arabs—their mental representations of Arabs are based entirely on media depictions, which themselves are notoriously dehumanizing (Shaheen, 2003; see also Esses, Medianu, & Lawson, 2013). There is at least some evidence (all exploratory) from our Qualtrics Panels sample of generators that is consistent with this possibility. For example, the median number of Arab friends that generators in our sample reported having was zero, making it plausible that many Americans' impressions of Arabs are shaped by (potentially negative) outside sources. Moreover, we found evidence suggesting that generators who *did* have more positive intergroup contact with Arabs harbored less dehumanizing representations of Arabs relative to Americans (see supplementals), suggesting that direct personal contact could help shape mental representations (and perhaps buffer against negative discourse). Still, these findings are only tentative, and they are well worth investigating more deeply in their own right.

Finally, future work should seek to identify what some of the boundary conditions of these findings might be. At the broadest level, it would be informative to know whether these findings are specific to intergroup dynamics in the United States (where Arab individuals tend to be heavily dehumanized), or whether they are instead representative of something broader about intergroup dehumanization. It is conceivable that the basic findings reported here reflect a broad psychological tendency for intergroup dehumanization to become internalized *even* by those who explicitly reject it. But this interpretation requires further empirical support, as our findings cannot (at this time) be generalized beyond U.S. samples. Another boundary condition worth examining concerns whether the findings reported here depend on which base face researchers

use in the reverse-correlation procedure. Our reverse-correlation procedure included a European base face that is commonly used with this technique (even in assessments of how East-Asian and North-African individuals are mentally represented: Dotsch & Todorov, 2012; Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Lundqvist & Litton, 1998). Still, it is possible that the shape and form that mental representations take could have been different had we used a less Eurocentric base image than the one used here (although there is less reason to believe that this would affect differences in the representations of high versus low explicit dehumanizers or right versus left-leaning generators). Lastly, these findings speak to only one kind of dehumanization that can pervade people's mental representations—that is, blatant, *animalistic* dehumanization. It would be informative for future researchers to examine whether these findings hold when the dependent variable concerns non-animalistic (e.g., mechanistic) dehumanization, instead.

In summary, even though liberals and conservatives self-report different beliefs about dehumanized outgroups (here, Arabs), their mental representations of those outgroups can be virtually identical—and indeed, unmistakably dehumanized. Moreover, the level of dehumanization in one's mental representations is not reducible—at least in our analysis—to obvious experimental confounds (like the fact that people tend to mentally represent outgroups more negatively than ingroups: Ratner et al., 2012). The implication of all this is that blatant dehumanization, which has been previously measured using self-reports alone, may be more widespread than previously thought. Still, the consequences of that dehumanization—how these representations shape human thought and behavior—are not well understood. Indeed, there is much work to be done with respect to understanding both the consequences and causes of harboring dehumanizing mental representations of outgroup members. We look forward to the

development of further research on this point, and indeed, to a deeper the understanding of blatant dehumanization more broadly.

Context of the Research

Our work to date has highlighted the prevalence and consequentiality of explicit forms of blatant dehumanization—the psychological tendency to overtly liken groups of people to non-human entities (Kteily & Bruenau, 2017b). That work suggested that a surprising proportion of individuals may willfully rate certain groups as being like 'lower' animals. But the majority of what we know about blatant dehumanization comes from studies that rely on self-report measures—measures that are notoriously subject to social desirability concerns (Gawronski & De Houwer, 2014). We also noticed across our research program that individuals on the political left were less likely to explicitly dehumanize low-status targets like Arabs than individuals on the political right. In combination, this made us curious whether we might find even broader evidence for blatant dehumanization—perhaps even among those on the political left—using less obtrusive measures. We relied here on the reverse-correlation image-classification procedure—an unobtrusive measure of blatant dehumanization which allows researchers to estimate participants' mental representations of target groups (Dotsch & Todorov, 2012). We found that even those who explicitly disavow dehumanization of Arabs nevertheless harbor mental representations of them that are heavily dehumanized, approaching the levels of explicit dehumanizers. Our results suggest that blatant dehumanization of low-status target groups may be much more widespread than self-report measures would have us believe. They also pose several questions for future research: Just how aware are people of the degree of dehumanization in their mental representations? And how much do these mental representations shape downstream behavior?

References

Altemeyer, B. (2007). *The authoritarians.* Winnipeg, Canada: University of Manitoba. Retrieved from http://home.cc.umanitoba.ca/altemey

Bar-Tal, D. (1989). Delegitimization: The extreme case of stereotyping and prejudice. In D. Bar-Tal, C. F. Graumann, A. W. Kruglanski, & W. Stroebe (Eds.), *Stereotyping and Prejudice: Springer Series in Social Psychology*. Springer, New York: NY.

Bastian, B., Denson, T. F., & Haslam, N. (2013). The roles of dehumanization and moral outrage in retributive justice. *PloS One, 8*, e61842.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48.

Bizumic, B., & Duckitt, J. (2018). Investigating right wing authoritarianism with a very short authoritarianism scale. *Journal of Social and Political Psychology*, *6*, 129-150.

Bodenhausen, G. V., & Petsko, C. D. (in press). Complications in predicting intergroup behavior from implicit biases: One size does not fit all. In J. A. Krosnick, T. H. Stark, & A. L. Scott (Eds.), *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge, UK: Cambridge University Press.

Bolger, N., Stadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285-301). New York, NY: Guilford Press.

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2017). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*, *28*, 92-103.

Bruneau, E., Jacoby, N., Kteily, N., & Saxe, R. (2018). Denying humanity: The distinct neural correlates of blatant dehumanization. *Journal of Experimental Psychology: General*, *147*, 1078-1093.

Bruneau, E., Kteily, N., & Laustsen, L. (2018). The unique effects of blatant dehumanization on attitudes and behavior towards Muslim refugees during the European 'refugee crisis' across four countries. *European Journal of Social Psychology*, *48*, 645-662.

Dotsch, R. (2016). Rcicr: Reverse-correlation image-classification toolbox. *R package version 0.3.4.1*. https://CRAN.R-project.org/package=rcicr

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*(5), 562-571.

Dotsch, R., Wigboldus, D. H. J., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*, 978-980.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Shabenberger, O. (2008). An R^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, *27*, 6137-6157.

Esses, V. M., Medianu, S., & Lawson, A. S. (2013). Uncertainty, threat, and the role of the media in promoting the dehumanization of immigrants and refugees. *Journal of Social Issues*, *69*, 518-536.

Esses, V. M., Veenvliet, S., Hodson, G., & Mihic, L. (2008). Justice, morality, and the dehumanization of refugees. *Social Justice Research*, *21*, 4-25.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (*2nd ed.)*. New York, NY: Cambridge University Press.

Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology, 94*, 292-306.

Goff, P. A., Jackson, M. C., Di Leone, B. A. L., Culotta, C. M., & DiTomasso, N. A. (2014). The essence of innocense: Consequences of dehumanizing Black children. *Journal of Personality and Social Psychology*, *106*, 526-545.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493-498.

Gunderson, A. B., & Kunst, J. R. (2019). Feminist ≠ feminine? Feminist women are visually masculinized whereas feminist men are feminized. *Sex Roles*, *80,* 291-309.

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., . . . Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO$_7$ scale. *Journal of Personality and Social Psychology, 109*(6), 1003-1028.

Hodson, G., & Costello, K. (2007). Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes. *Psychological Science*, *18*, 691-698.

Jackson, L. E., & Gaertner, L. (2010). Mechanisms of moral disengagement and their differential use by right-wing authoritarianism and social dominance orientation in support of war. *Aggressive Behavior*, *36*, 238-250.

Jackson, J. C., Hester, N., & Gray, K. (2018). The faces of God in America: Revealing religious diversity across people and politics. *PLoS ONE*, *13*, e0198745.

Kroh, M. (2007). Measuring left-right political orientation: The choice of response format. *Public Opinion Quarterly*, *71*, 204-220.

Kteily, N., & Bruneau, E. (2017). Backlash: The politics of real-world consequences of minority group dehumanization. *Personality and Social Psychology Bulletin*, *43*, 87-104.

Kteily, N. S., & Bruneau, E. (2017). Darker demons of our nature: The need to (re)focus attention on blatant forms of dehumanization. *Current Directions in Psychological Science*, *26*, 487-494.

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of Personality and Social Psychology, 109*(5), 901-931.

Kunst, J. R., Kteily, N., & Thomsen, L. (2018). "You little creep": Evidence of blatant dehumanization of short groups. *Social Psychological and Personality Science*, *10*, 160-171.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*, 569-585.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models. *R package version 2.0-33*. https://CRAN.R-project.org/package=lmerTest

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355-362.

Lei, R. F., & Bodenhausen, G. V. (2017). Racial assumptions color the mental representation of social class. *Frontiers in Psychology*, *8*, 519. doi:10.3389/fpsyg.2017.00519

Leyens, J-P., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review*, *4*, 186-197.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442.

Loughnan, S., & Haslam, N. (2007). Animals and androids: Implicit associations between social categories and non-humans. *Psychological Science*, *18*, 116-121.

Loughnan, S., Haslam, N., & Kashima, Y. (2009). Understanding the relationship between attribute-based and metaphor-based dehumanization. *Group Processes & Intergroup Relations*, *12*, 747-762.

Lundqvist, D. & Litton, J. E. (1998). The Averaged Karolinska Directed Emotional Faces – AKDEF. CD ROM from department of clinical neuroscience, psychology section, Karolinska Institutet (Tech. Rep.). ISBN 91-630-7164-9.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*(2), 209-226.

Maoz, I., & McCauley, C. (2008). Threat, dehumanization, and support for retaliatory aggressive policies in asymmetric conflict. *Journal of Conflict Resolution*, *52*, 93-116.

McClelland, G. H., Lynch, J. G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false-positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology*, *25*, 679-689.

Monteith, M. J., & Mark, A. Y. (2005). Changing one's prejudiced ways: Awareness, affect, and self-regulation. *European Review of Social Psychology*, *16*, 113-154.

Nosek, B. A., Banaji, M. R., & Jost, J. T. (2009). The politics of intergroup attitudes. In J. T. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Attitudes: Insights from the New Implicit Measures* (pp.65-82). Hillsdale, NJ: Erlbaum.

Page-Gould, E., Sharples, A. E., & Song, S. (2019, October). *Effect sizes for models of longitudinal data*. In P. Shrout (Chair), Modeling Mediation Processes in Longitudinal Data. Symposium conducted at the annual meeting of the Society of Experimental Social Psychology, Toronto, ON.

Petsko, C. D. (2020, June 10). Mental Representations of Arabs. https://doi.org/10.17605/OSF.IO/5MWPQ

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811-832.

Ratner, K. G., Dotsch, R., Wigboldus, D. H., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology, 106*, 897-911.

Roberts, S. O., Weisman, K., Lane, J. D., Williams, A., Camp, N., Wang, M., Robison, M., Sanchez, K., Griffiths, C. (2020). God as a White man: A psychological barrier to conceptualizing Black people and women as leadership worthy. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037/pspi0000233

Saminaden, A., Loughnan, S., & Haslam, N. (2010). Afterimages of savages: Implicit associations between 'primitives', animals and children. *British Journal of Social Psychology*, *49*, 91-105.

Shaheen, J. G. (2003). Reel bad Arabs: How Hollywood vilifies people. *The Annals of the American Academy*, *588*, 171-193.

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G., & McClelland, G. (2012). Spotlights, floodlights, and the magic number zero: Simple effect tests in moderated regression. *Journal of Marketing Research*, *L*, 277-288.

Viki, G. T., Osgood, D., & Phillips, S. (2013). Dehumanization and self-reported proclivity to torture prisoners of war. *Journal of Experimental Social Psychology*, *49*, 325-328.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*, 383-388.