Attribution of undesirable character traits, rather than trait-based dehumanization, predicts punishment decisions

Robert A. Brennan[1], Florence E. Enock[1,2], & Harriet Over[1]

[1]Department of Psychology, University of York, York, UK. YO10 5DD
[2]The Alan Turing Institute, British Library, 96 Euston Road, London, UK. NW1 2DB

Address for correspondence: Robert A. Brennan, rb1733@york.ac.uk; brennan.rob.a@gmail.com

1 Abstract

Previous work has reported that the extent to which participants dehumanized criminals by denying them uniquely human character traits such as refinement, rationality and morality predicted the severity of the punishment endorsed for them. We revisit this influential finding across six highly powered and pre-registered studies. First, we conceptually replicate the effect reported in previous work, demonstrating that our method is sensitive to detecting relationships between trait-based dehumanization and punishment should they occur. We then investigate whether the apparent relationship between trait-based dehumanization and punishment is driven by the desirability of the traits incorporated into the stimulus set, their perceived humanness, or both. To do this, we asked participants to rate the extent to which criminals possessed uniquely human traits that were either socially desirable (e.g. cultured and civilized) or socially undesirable (e.g. arrogant and bitter). Correlational and experimental evidence converge on the conclusion that apparent evidence for the relationship between trait-based dehumanization and punishment is better explained by the extent to which participants attribute socially desirable attributes to criminals rather than the extent to which they attribute uniquely human attributes. These studies cast doubt on the hypothesized causal relationship between trait-based dehumanization and harm, at least in this context.

Keywords: dehumanization, harm, intergroup bias, social cognition, trait attribution

2 Introduction

Understanding the motivations that lead to intergroup harm has been a driving force behind social psychology since its conception (Farr, 1996; Gaines & Reed, 1995). Many psychological processes have been shown to exacerbate outgroup derogation, including prejudicial attitudes and stereotyping (Dovidio et al., 2010; Major & Sawyer, 2009). Over the last 25 years, an increasing body of research has investigated the extent to which a psychological process of dehumanization increases the risk of intergroup harm (Bloom, 2017, 2022; Giner-Sorolla et al., 2021; Goldenberg et al., 2021; Lang, 2010, 2020; Manne, 2016, 2018; Over, 2021b, 2021a; Rai et al., 2017; Ruiter, 2022; Smith, 2011, 2020, 2021; Vaes et al., 2021). According to the dehumanization hypothesis, when members of an outgroup are perceived as less human than ingroup members, they are at greater risk of harm (Haslam, 2019; Haslam & Loughnan, 2014, 2016; Over, 2021b; Smith, 2011, 2016). Subtle forms of dehumanization are thought to be pervasive in contemporary society. For example, the dehumanization of national groups (Bain et al., 2011; Leyens et al., 2003), religious groups (Banton et al., 2020; Leidner et al., 2013), individuals on a low income (Loughnan et al., 2013), and refugees (Bruneau et al., 2018; Esses et al., 2008) has been reported in the literature.

Within social psychology, several different characterisations of dehumanization have been proposed. Infrahumanization theory (Leyens et al., 2000, 2001) posits that a subtle form of dehumanization occurs where

outgroup members are viewed as experiencing uniquely human emotions such as pride and melancholy to a lesser extent than ingroup members. The mental state account maintains that outgroup members are dehumanized to the extent that they are denied mental states (Gray et al., 2007; Hare & Woods, 2020; Harris & Fiske, 2006).

The dual model of dehumanization is of particular interest to the current research (Haslam, 2006). According to the dual model, individuals and groups are dehumanized to the extent that they are denied uniquely human character traits. The dual model distinguishes between two forms of dehumanization (Haslam et al., 2004). When outgroup members are *animalistically dehumanized*, or perceived as similar to animals, they are thought to possess traits such as civility, refinement, rationality, moral sensibility, and maturity to a lesser extent than the ingroup. When outgroup members are *mechanistically dehumanized*, or perceived as similar to robots, they are thought to possess traits such as emotional responsiveness, interpersonal warmth, depth, cognitive openness, and agency to a lesser extent than the ingroup.

According to the dual model, the more an individual or group is either animalistically or mechanistically dehumanized, the greater their risk of being harmed (Haslam, 2006; Haslam & Loughnan, 2014). Haslam & Loughnan (2014) argue that "*dehumanization is important as a psychological phenomenon because it can be so common and yet so dire in its consequences*" (p. 401). Haslam (2021) further notes that "*Many studies have examined how dehumanizing perceptions enable harm or provide support for it. Some of this work points to direct links between tendencies to dehumanize others and... aggressive behaviour*" (pp. 139). Empirical research has suggested that trait-based dehumanization facilitates social exclusion (Bastian & Haslam, 2010) and reduces prosocial behaviour (Andrighetto et al., 2014).

Bastian et al. (2013) conducted influential empirical studies testing the hypothesised association between the denial of human character traits and the endorsement of harsh punishment (see also Barber & Davis (2022); Chen-Xia et al. (2023); Kasper et al. (2022); Morehouse et al. (2023); Rousseau et al. (2023); West & Thomson (2022)). The researchers measured how trait-based dehumanization influenced participants' punishment of criminals. Participants were asked to rate their agreement with four items assessing animalistic dehumanization of criminals: "*I felt like the person in the story was refined and cultured*" [reversed], "*I felt like the person in the story was rational and logical, like they were intelligent*" [reversed] "*I felt like the person in the story lacked self-restraint, like an animal*", and "*I felt like the person in the story was unsophisticated*". Participants were also asked to rate their agreement with four items assessing mechanistic dehumanization of criminals: "*I felt like the person in the story was open minded, like they could think clearly about things*" [reversed], "*I felt like the person in the story was emotional, like they were responsive and warm*" [reversed], "*I felt like the person in the story was superficial like they had no depth*", "*I felt like the person in the story was mechanical and cold, like a robot*". Bastian and colleagues reported that both forms of dehumanization predicted endorsement of harsh punishment for the criminals portrayed in their stimuli, concluding that their participants viewed criminals as '*subhuman and beastly*' (pp.9).

Recently, however, the explanatory value of the dual model has been called into question (Enock, Tipper, et al., 2021; Over, 2021b, 2021a). According to these critiques, evidence for trait-based dehumanization is often confounded with social desirability. In Bastian and colleagues' work, evidence that criminals were animalistically dehumanized was drawn from the observation that participants judged them to be unsophisticated, lacking self-restraint, unrefined, uncultured, irrational, and unintelligent. Evidence that criminals were mechanistically dehumanized came from the observation that participants viewed them as superficial, cold, and lacking in warmth and responsiveness. These results may reflect dehumanization because the traits criminals were found to lack are those perceived as uniquely or essentially human (Haslam et al., 2004; Haslam, 2006). However, as the traits deemed uniquely humans were all socially desirable, evidence for trait-based dehumanization cannot be separated from evidence of negative evaluation more generally. An alternative explanation for the findings from Bastian and colleagues is that participants endorse harsh punishment against criminals to the extent they perceive criminals to possess undesirable or antisocial characteristics.

Bastian et al. (2013) seek to account for this possibility by statistically controlling for participants' moral outrage at the targets' behaviour in their analysis. They report that the relationship between trait-based dehumanization and punishment remains even when moral outrage is controlled for. While this is interesting and suggestive of the independent effects of dehumanization, it cannot fully address the conceptual

weaknesses in how dehumanization was operationalised. A more convincing way to de-confound evidence for trait-based dehumanization from evidence of negative evaluation is to ask participants to rate the target group on traits that are uniquely human but vary from socially desirable to undesirable Over (2021a). Previous research conducted by Enock, Flavell, et al. (2021) has established that undesirable character traits such as jealous, spiteful and bitter are considered unique to humans and socially undesirable. Across three intergroup contexts, the researchers found that participants attributed socially desirable human traits more strongly to the ingroup and socially undesirable traits more strongly to the outgroup (see also Decker & Lord (2023); Enock, Tipper, et al. (2021); Enock & Over (2022); Enock & Over (2023)). Enock, Flavell, et al. (2021) concluded that intergroup preference may better explain apparent evidence for trait-based dehumanization. However, it is not yet clear how the attribution of uniquely human character traits relates to harm. Addressing this question is crucial to understanding the extent to which the dual model of dehumanization can help explain real-world discrimination and negativity.

We revisit the hypothesised causal relationship between trait-based dehumanization and harm in the context of endorsing harsh punishment for criminals. In Studies 1A and 1B, we seek to conceptually replicate the key findings of Bastian and colleagues (2013), suggesting that the extent to which participants animalistically (Study 1A) and mechanistically (Study 1B) dehumanize criminals predicts the severity of the punishment participants endorse for them. In Studies 2A and 2B, we adopt a similar design but incorporate socially undesirable traits into our stimulus set. This addition to the design allows us to investigate whether trait-based dehumanization, undesirable trait attribution, or both predict the severity of punishment. Following Bastian and colleagues, and to understand the generalisability of our findings, we investigate these questions in relation to two different types of crime (violent crime and theft). In Studies 3A and 3B, we seek to investigate a similar question using an experimental design. We present participants with vignettes in which criminals are described using character traits that differ in how socially desirable they are and whether or not they are unique to humans. We also measure how these varying descriptions influence participants' parole decisions. This allows us to directly measure whether there is a causal relationship between trait-based dehumanization and punishment, independent of ingroup preference.

3 Methods

All studies received ethical approval from the Psychology Departmental Ethics Committee at the University of York (approval number 926). All data collection occurred online, and the studies were created and administered using Qualtrics (https://www.qualtrics.com). Participants were recruited through the online platform Prolific (https://www.prolific.co), with an independent sample recruited for each study. Informed consent was obtained at the start of each session according to approved ethical guidelines. Inclusion criteria for each study included adult participants fluent in English who had never been to prison for committing a crime and had a Prolific approval rating of at least 90% (95% for Studies 3A and 3B). Increases in Prolific's recommended rate of compensation for participation during data collection meant the reward ranged from approximately £7 per hour in Studies 1A and 1B to approximately £8 in the other four studies. Assumption testing and analyses were conducted using *SPSS* and *RStudio*. All studies were pre-registered on AsPredicted.com before commencing data collection. Links to pre-registration documents, data files (Brennan et al., n.d.), a fully computationally reproducible version of the manuscript, and electronic supplementary materials including the stimuli used for each study can be found at: http://doi.org/10.17605/OSF.IO/D4CVP.

4 Study 1A

Bastian et al. (2013) presented evidence that the more participants dehumanize violent criminals, the harsher the punishment participants endorse for them. We sought to test whether we could conceptually replicate this relationship between trait-based dehumanization and punishment using terms very similar to those used by Bastian and colleagues. In study 1A, participants read a series of scenarios describing fictitious criminals and their violent crimes. Following this, participants rated the extent to which the criminals possessed four character traits that distinguish humans from animals (*refined, rational and logical, has a sense of*

morality and *civilised*). Following common practice in the field, we refer to these as uniquely human traits. Participants also rated the extent to which criminals possessed four character traits that distinguish humans from machines (*openminded, emotionally responsive, has a depth of character*, and *interpersonally warm*). Following common practice in the field, we refer to these as human nature traits. Participants also responded to an item measuring how harsh they thought the violent criminals' punishment should be. Following Bastian et al. (2013), we predicted that the less participants attributed these uniquely human traits to criminals, the harsher the punishment they would recommend for criminals.

4.1 Participants

A power analysis using G*Power indicated that a sample size of 89 would allow us to detect a medium effect size ($f^2 = 0.15$) with an alpha of 0.05 and power of 0.95. A final sample of 100 participants was collected, with 54 identifying as female, 44 as male and 2 as non-binary. Ages ranged from 18 to 63 ($M = 26.53$, $SD = 8.94$). In accordance with our pre-registered exclusion criteria, data submitted by six individuals who failed one or both attention checks (i.e., gave a response more than 20 points away from the instructed end of the scale) were omitted and replaced. Participation took an average of approximately eight minutes.

4.2 Materials

Vignettes. All participants responded to the same five vignettes detailing different scenarios involving violent crimes. An effort was made to ensure that all five vignettes were similar in length, degree of detail, and severity of crimes depicted. In each vignette, the target's name and pronouns were gender-neutral, their age and ethnicity were not indicated, and the scenarios depicted were all set in unspecified locations. All vignettes are included in the Supplementary materials. For example: "*Charlie was arrested after a fight broke out in a pub soon after opening time, apparently triggered by a minor disagreement. Charlie smashed a pint glass and used it to stab another customer. Two additional customers received cuts as they tried to hold Charlie back until the Police arrived.*"

Trait attribution. After reading each vignette, participants responded to items designed to measure trait-based dehumanization, broadly following the procedure of Bastian and colleagues (2013). Participants indicated the extent to which they attributed four uniquely human traits (*refined, rational and logical, has a sense of morality* and *civilised*) and four human nature traits (*openminded, emotionally responsive, has a depth of character*, and *interpersonally warm*) to the criminals depicted. Participants indicated their agreement with each item (e.g., '*I think* [e.g., *Charlie*] *is refined*') using an unmarked sliding scale from 0 ('*Strongly Disagree*') to 100 ('*Strongly Agree*'), with the sliders initially fixed at the midpoint. According to the dual model, lower scores indicated greater dehumanization of violent criminals. An attention check appeared halfway through the dehumanization items for two criminals ('*Please move the slider all the way to Strongly Agree/Disagree*').

Harshness of punishment endorsed. Using an unmarked sliding scale that ranged from 0 ('*Not at all harsh*') to 100 ('*Very harsh*'), participants were asked to respond to the question '*How harsh do you think the punishment for* [e.g., *Charlie*] *should be*?'.

4.3 Design

Following Bastian et al. (2013), we utilised a within-subjects, correlational design. All participants read the same five vignettes presented in a random order and responded to the same trait attribution and punishment items. Participants' scores for the trait attribution items and the endorsed harshness of punishment item were then averaged across scenarios. The presentation of the items in the trait attribution task was also randomised.

4.4 Procedure

Participants were informed that the study would examine how social attributions influence our behavioural intentions towards criminals. After providing informed consent, participants answered a few demographic questions and confirmed that they had never been to prison for committing a crime. The first of five vignettes

then followed. After reading the vignette, participants were asked to respond to the trait attribution items, followed by the single item asking them to indicate how harshly they thought the criminal should be punished. Participants repeated the above steps for each of the remaining four vignettes. To ensure participants read the stimuli carefully, each vignette remained on the screen for at least 15 seconds.

4.5 Results

4.5.1 Model 1: Animalistic dehumanization and punishment

In line with our pre-registered criteria, this analysis omitted two highly influential cases (remaining sample N = 98). We first calculated the average attribution score for uniquely human traits and punishment for each participant in the sample. We then conducted a simple linear regression to understand whether the extent to which participants attributed uniquely human traits to criminals predicted the harshness of punishment participants endorsed for them. A significant negative relationship was found, b = -0.56 [-0.75, -0.37], t = -5.93, p <0.001, see Figure 1. Thus, the more violent criminals were animalistically dehumanized (by being denied uniquely human traits), the harsher the punishment participants endorsed. The model explained approximately 27% of the variance in the harshness of punishment scores, R^2 = 0.27, $F(1, 96)$ = 35.14.

4.5.2 Model 2: Mechanistic dehumanization and punishment

In line with our pre-registered criteria, seven highly influential were omitted from the analysis (remaining sample N = 93). After calculating the average attribution score for human nature traits and punishment for each participant, we conducted a simple linear regression to test attribution of human nature traits predicted the harshness of punishment endorsed for violent criminals. A significant negative relationship was found, b = -0.41 [-0.57, -0.24], t = -4.93, p <0.001, see Figure 1. This relationship shows that greater mechanistic dehumanization (operationalised as the denial of human nature traits) was associated with the endorsement of harsher punishment. The model explains 21% of the variance in the harshness of punishment scores, R^2 = 0.21, $F(1, 91)$ = 24.29.

5 Study 1B

Study 1B investigates whether the relationship found in Study 1A replicates when participants are asked to judge a different type of criminal activity. In Study 1B, we examined whether animalistic and mechanistic dehumanization, as operationalised by Bastian et al. (2013), are associated with the harshness of punishment endorsed for individuals who commit theft. The design, materials and analysis plan were similar to that used in Study 1A, except that the scenarios involved theft rather than violent crime.

5.1 Method

5.1.1 Participants

Based on the same power analysis used in Study 1A, a sample of 100 participants was collected, with 55 identifying as male and 45 as female. Ages ranged from 18 to 57 (M = 25.45, SD = 8.25). The attention checks in Study 1A were also used in Study 1B. Ten participants failed one or both attention checks, and their data was omitted and replaced as per our pre-registration. Participation took an average of nine minutes.

5.1.2 Materials

The measures of dehumanization and punishment were identical to those used in Study 1A.

Vignettes All participants responded to the same five vignettes, each detailing a crime involving theft (see Supplementary materials). As in Study 1A, an effort was made to ensure the vignettes were similar in structure and amount of detail. Once again, all of the perpetrators had gender-neutral names. An example of one of the theft vignettes is as follows: "*Until their recent arrest, Charlie had worked as a till operator at a local charity shop supporting individuals experiencing homelessness. Charlie had been stealing cash amounts*

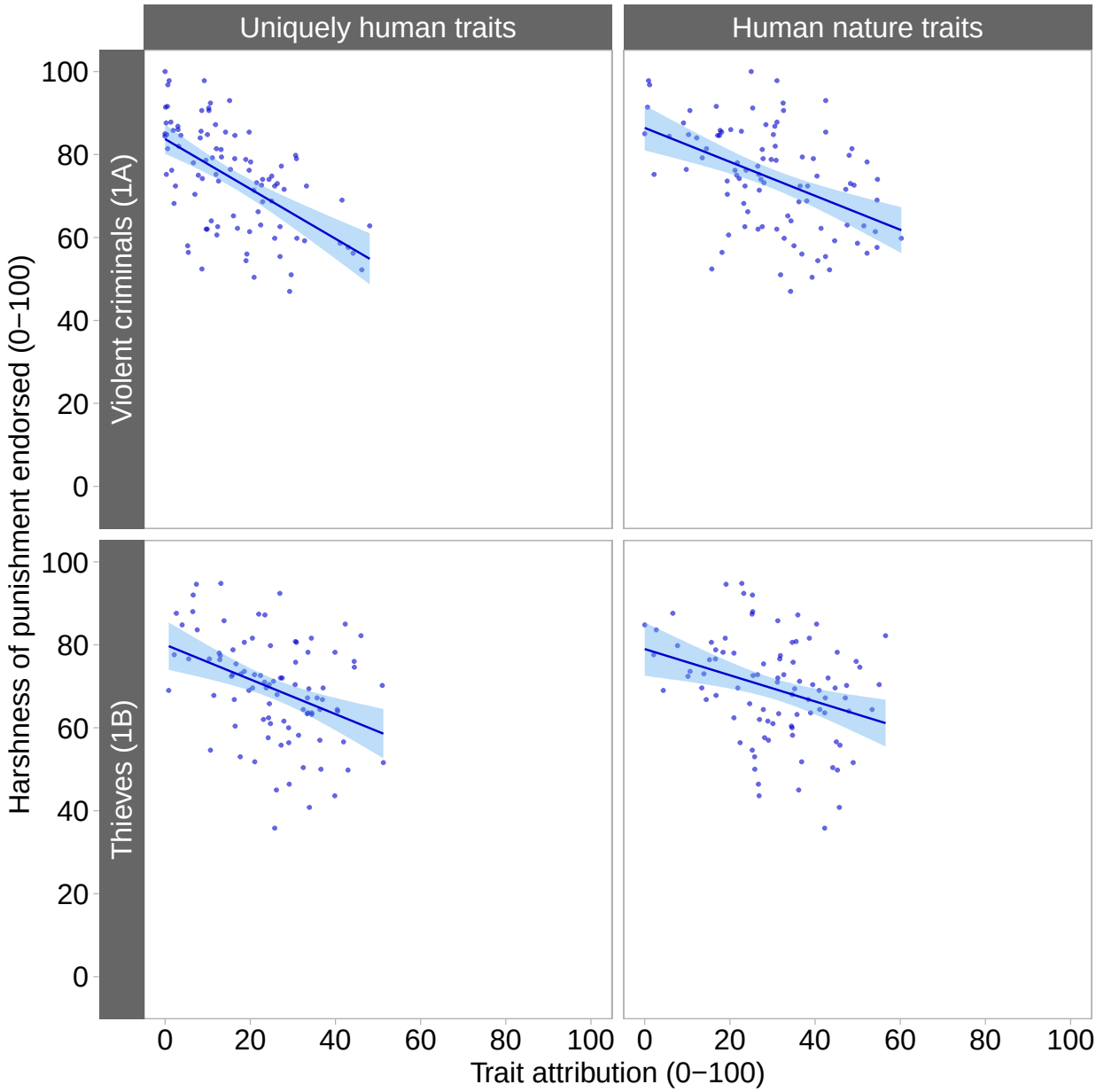

Figure 1: Results of Study 1. Seemingly in line with Bastian and colleagues (2013), greater animalistic (left) and mechanistic (right) dehumanization of violent criminals (Study 1A, top) and thieves (Study 1B, bottom) was associated with harsher punishment.

varying from £5 to £50 from the tills almost daily over a five-year period. Police revealed that Charlie had stolen several thousand pounds from the charity shop while working there."

5.2 Results

5.2.1 Model 1: Animalistic dehumanization and punishment

Seven highly influential cases were omitted from the analysis (remaining sample $N = 93$). We calculated the average attribution scores for uniquely human trait attribution and punishment for each participant and then conducted a simple linear regression to measure whether trait attribution predicted the harshness of punishment endorsed for thieves. As shown in Figure 1, a significant negative relationship was found, b =-0.42 [-0.64, -0.21], t = -3.97, p <0.001. Thus, greater animalistic dehumanization of thieves was associated with the endorsement of harsher punishment for them. The model explains approximately 15% of the variance in the harshness of punishment scores, $R^2 = 0.15$, $F(1, 91) = 15.75$.

5.2.2 Model 2: Mechanistic dehumanization and punishment

Eight highly influential cases were omitted from the analysis (remaining sample $N = 92$). After calculating the average score of human trait attribution and punishment, we conducted a simple linear regression to test whether or not human trait attribution predicted the harshness of punishment endorsed for thieves. A significant negative relationship was found, b = -0.27 [-0.47, -0.08], t = -2.77, p = 0.007. These data show that greater mechanistic dehumanization is associated with the endorsement of harsher punishment for thieves (see Figure 1). The model explains approximately 8% of the variance in the harshness of punishment scores, $R^2 = 0.08$, $F(1, 90) = 7.65$.

6 Study 2A

Study 2A investigated whether apparent evidence for a relationship between trait-based dehumanization and endorsement of harsh punishment for violent criminals remains when controlling for the desirability of the traits. We tested this by introducing character traits perceived as uniquely human yet socially undesirable into the stimulus set (Enock, Tipper, et al., 2021; Over, 2021b). The dual model predicts that to the extent criminals are denied uniquely human character traits, they will be subjected to harsher punishment. We predict that trait desirability will moderate the relationship between human trait attribution and punishment. More specifically, we predict that the extent to which violent criminals are denied socially desirable character traits, and attributed socially undesirable character traits, will predict harsh punishment.

6.1 Method

6.1.1 Participants

A power analysis using G*Power indicated that a sample size of 119 would allow us to detect a medium effect size ($f^2 = 0.15$), with three predictors (trait attribution; trait desirability; attribution*desirability), an alpha of 0.05 and power of 0.95. To counterbalance the sample equally and allow for the exclusion of outliers, a sample of 130 was collected. Within the sample, 66 identified as female, 62 as male, and 2 as non-binary. Ages ranged from 18 to 55 ($M = 28.46$, $SD = 9.14$). Similar to Studies 1A and 1B, two attention checks were included in this study. As per our preregistered plan, 16 participants failed one or both attention checks; thus, their data were omitted and replaced. Participation took an average of nearly eight minutes.

6.1.2 Design

This study utilised a mixed design. All participants responded to items designed to measure animalistic dehumanization and mechanistic dehumanization. The desirability of the traits rated by participants was manipulated between subjects: half of the participants rated criminals on the extent to which they possessed socially desirable traits, and half rated criminals on the extent to which they possessed undesirable traits. All participants responded to a single item measuring the harshness of punishment endorsed.

6.1.3 Materials

Vignettes. All participants read the same vignettes describing violent crimes as in Study 1A.

Trait attribution. After reading each vignette, participants responded to an 8-item scale measuring animalistic dehumanization (4 items) and mechanistic dehumanization (4 items) of the criminal portrayed. Participants made trait attributions by indicating their agreement with each item using an unmarked sliding scale ranging from 0 ('*Strongly Disagree*') to 100 ('*Strongly Agree*'), all of which were initially positioned at the scale's midpoint. Depending on the condition, the eight trait items were either socially desirable (*uniquely human: cultured, civilised, sophisticated, moral; human nature: generous, open-minded, warm, kind*) or socially undesirable (*uniquely human: corrupt, controlling, arrogant, bitter; human nature: jealous selfish, spiteful, cruel*). The lower the score, the more participants dehumanize the criminal target by denying them human traits.

Harshness of punishment endorsed. The same single-item scale for measuring the harshness of punishment endorsed in Studies 1A and 1B was employed in Study 2A.

6.1.4 Procedure

The procedure in Study 2A mirrored that of Studies 1A and 1B.

6.2 Results

6.2.1 Model 1: Animalistic dehumanization and punishment

Eight highly influential cases were omitted from the analysis (remaining sample $N = 122$). The regression model tested for a relationship between participants' average scores for uniquely human trait attribution and harshness of punishment endorsed with trait desirability included as a moderator (desirable = 0, undesirable = 1).

The moderated regression showed no significant effect of uniquely human trait attribution on punishment, b = -0.07 [-0.22, 0.09], t = -0.83, p = 0.408. Thus, when undesirable uniquely human traits were included in the measure of animalistic dehumanization, the previously reported relationship between animalistic dehumanization and the endorsement of harsher punishment (Bastian et al., 2013) was no longer significant.

The interaction between uniquely human trait attribution and trait desirability was significant, b = 1.31 [1, 1.63], t = 8.3, p = <0.001. In line with our prediction, simple slopes showed that the more socially desirable human traits participants attributed to criminals, the less harshly participants thought they should be punished, b = 0.58, [0.39, 0.78], t = 5.9, p = <0.001. The more undesirable traits participants attributed to criminals, the more harshly participants thought they should be punished, b = -0.73, [-0.98, -0.49], t = -5.92, p = <0.001 (see Figure 2). The model explained approximately 38% of the variance in the harshness of punishment endorsed, R^2 = 0.38, $F(3, 118)$ = 23.61.

6.2.2 Model 2: Mechanistic dehumanization and punishment

Eight highly influential cases were omitted from the analysis (remaining sample $N = 122$). A moderated regression analysis tested for a relationship between the average scores of human trait attribution and harshness of punishment endorsed to violent criminals and whether this interacted with trait desirability.

The moderated regression showed no significant effects of human nature trait attribution on punishment, b = -0.04 [-0.19, 0.12], t = -0.44, p = 0.658. The effect reported by Bastian and colleagues (2013), whereby mechanistic dehumanization predicted harsher punishment endorsement, which we replicated in Studies 1A and 1B, did not appear when undesirable human nature traits were included in our measures.

The interaction between uniquely human trait attribution and trait desirability was significant, b = 1.37 [1.06, 1.68], t = 8.65, p = <0.001. Simple slopes indicated that the more participants attributed socially desirable traits to criminals, the less harshly they thought those criminals should be punished b = -0.73, [-0.96, -0.51], t = -6.46, p = <0.001. As shown in Figure 2, the more participants attributed socially undesirable traits to criminals, the more harshly they thought those criminals should be punished, b = 0.64, [0.42, 0.86], t =

5.77, $p = <0.001$. The model explained 39% of the variance in the harshness of punishment scores, $R^2 = 0.39$, $F(3, 118) = 25.05$.

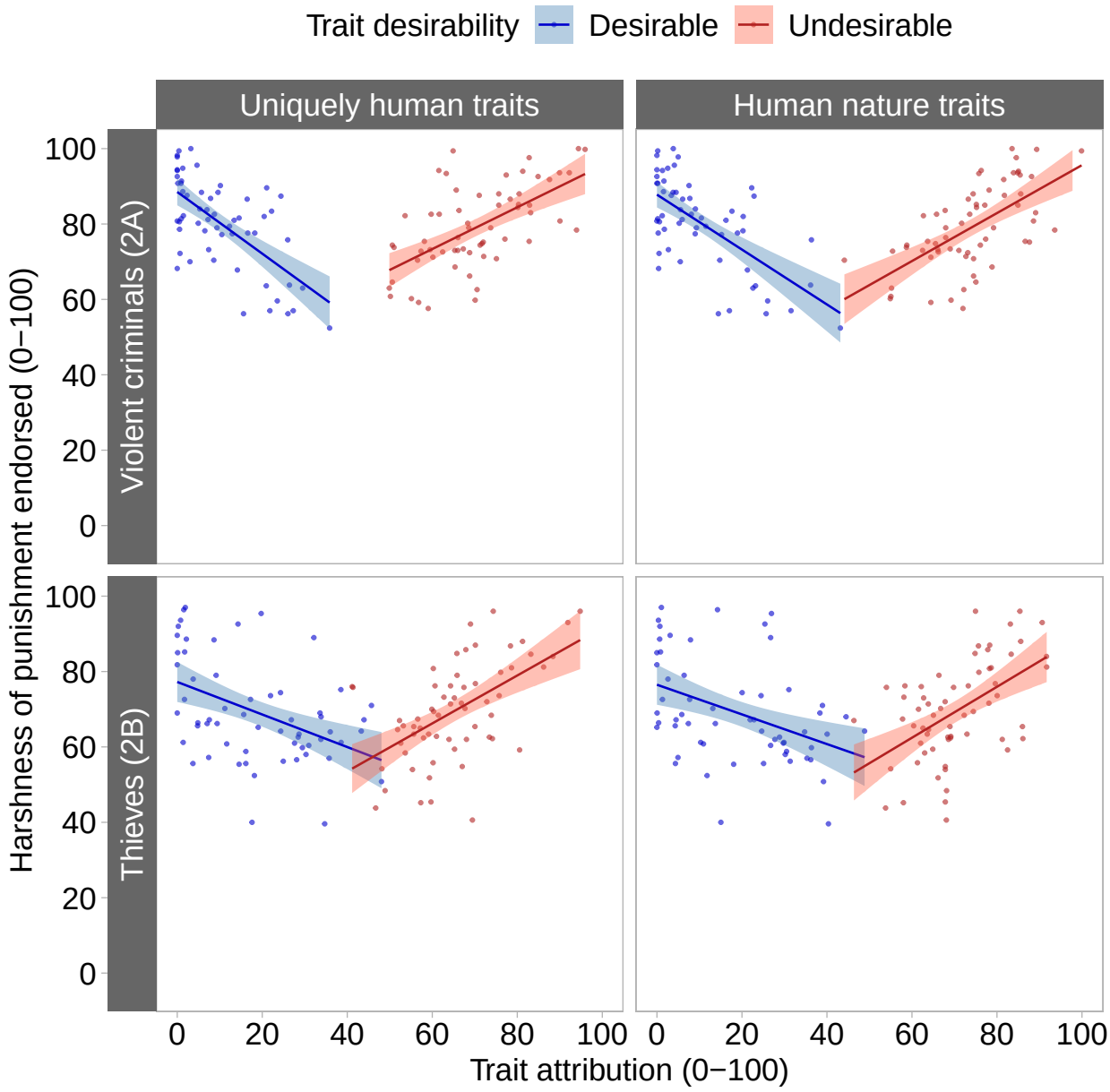

Figure 2: Results of Study 2: The relationship between trait attribution and punishment for violent criminals (Study 2A, top) and thieves (Study 2B, bottom) depends on the social desirability of the traits.

7 Study 2B

Study 2B sought to replicate the results of Study 2A but with thieves as the target group rather than violent criminals. We examined whether the apparent relationship between trait-based dehumanization and the endorsement of harsh punishment for thieves is better explained by the desirability of the traits incorporated into the stimulus set. We investigate this question using a very similar design and procedure to Study 2A, with the exception that the vignettes are those used in Study 1B detailing crimes involving theft. As in Study

2A, we hypothesise that trait desirability will moderate the relationship between human trait attribution and punishment. More specifically, we predict the extent to which criminals are denied socially desirable character traits, and attributed socially undesirable character traits, will predict endorsement of harsher punishment.

7.1 Method

7.1.1 Participants

The power analysis described in Study 2A informed the sample size for Study 2B. A separate sample of 130 participants was collected, of whom 74 identified as male, 53 as female, and three as non-binary. Ages ranged from 18 to 59 (M= 26.5, SD = 7.48). Data submitted by 20 participants who did not pass one or both checks were omitted and replaced. Participation took an average of nine and a half minutes.

7.1.2 Design

This study utilised a mixed-methods design, matching that of Study 2A. The same attention checks used in Studies 1A, 1B and 2A were used in Study 2B.

7.1.3 Materials

Vignettes. All participants responded to the same five vignettes used in Study 1B detailing scenarios involving criminals committing theft.

Trait attribution. The same scales for measuring animalistic dehumanization, mechanistic dehumanization, and punishment used in Study 2A were used in Study 2B.

7.1.4 Procedure

The procedure in Study 2B was identical to that of Study 2A, except for the vignettes describing crimes involving theft rather than violence.

7.2 Results

7.2.1 Model 1: Animalistic dehumanization and punishment

Eight highly influential cases were omitted from the analysis (remaining sample N = 122). A moderated regression tested for a relationship between average scores of uniquely human traits and harshness of punishment endorsed and whether this interacted with trait desirability. The moderated regression showed no significant effects of uniquely human trait attribution on punishment $b = 0.13$ [-0.05, 0.3], $t = 1.44$, $p = 0.152$. Replicating the results of Study 2A, when socially undesirable traits were incorporated into the stimulus set, there was no longer any relationship between trait-based dehumanization and punishment.

The interaction between uniquely human trait attribution and trait desirability was significant, b = 1.08 [0.74, 1.42], t = 6.21, p = <0.001. As illustrated in Figure 2, the more participants attributed socially desirable traits to criminals, the less harshly participants felt they should be punished, b = -0.43, [-0.65, -0.21], t = -3.88, p = <0.001. The more participants attributed undesirable traits to criminals, the more harshly participants felt they should be punished, b = 0.65, [0.38, 0.91], t = 4.85, p = <0.001. The model explained about 25% of the variance in endorsed harshness of punishment scores, R^2 = 0.25, $F(3, 118)$ = 12.88.

7.2.2 Model 2: Mechanistic dehumanization and punishment

The analysis omitted six highly influential cases (remaining sample N = 124). A regression tested for a relationship between human trait attribution and punishment and whether this interacted with trait desirability. As in Study 2A, and contradicting the findings of Bastian and colleagues (2013), the moderated

regression showed no significant relationship between human trait attribution and punishment, $b = 0.15$ [-0.04, 0.34], $t = 1.57$, $p = 0.119$.

However, the interaction between human nature trait attribution and trait desirability was significant, $b = 0.95$ [0.57, 1.33], $t = 5$, $p = <0.001$. As can be seen in Figure 2, simple slopes showed that the more participants attributed socially desirable human traits to criminals, the less harshly participants thought they should be punished, $b = -0.33$, [-0.55, -0.11], $t = -2.99$, $p = 0.003$. The more participants attributed socially undesirable human traits to criminals, the more harshly participants thought they should be punished, $b = 0.62$, [0.31, 0.92], $t = 4.01$, $p = <0.001$. The model explained about 17% of the variance in endorsed harshness of punishment scores, $R^2 = 0.17$, $F(3, 120) = 8.44$. These data suggest that the apparent relationships between animalistic and mechanistic dehumanization and punishment reported in previous research (Bastian et al., 2013) are better explained by the social desirability of the traits.

8 Study 3A

In Study 3A, we used an experimental design to examine further the hypothesised causal relationship between trait-based dehumanization and punishment when controlling for the social desirability of human traits incorporated into the stimuli. We described criminals with traits that varied in desirability and perceived humanness creating a 2×2 design. We then measured participants' willingness to endorse parole for each criminal described. The dual model predicts that criminals who are described in uniquely human terms will be more likely to be granted parole. We predicted that criminals described in socially desirable terms will be more likely to be granted parole. In principle, this design allows us to detect independent effects of dehumanization and trait sociability or an interaction between the two. In Study 3A, we specifically measure the extent to which animalistic dehumanization is causally related to parole decisions. Thus, we included uniquely human traits and those shared with other animals in our measures. We predict that participants will be more likely to endorse parole for criminals described with socially desirable traits, regardless of whether or not those traits are uniquely human or shared with other animals.

8.1 Methods

8.1.1 Participants

A power analysis using G*Power, with effect size specification as in SPSS, indicated that a sample size of 135 would allow us to detect a medium effect size ($\eta_p^2 = .09$) with a 2×2 factorial, repeated measures design, an alpha of .05, and a power of .95. To counterbalance the sample equally, a sample of 136 participants was collected, of whom 78 identified as female, 55 as male, 2 as non-binary, and 1 who preferred not to indicate their gender identity. Ages ranged from 18 to 63 ($M = 24.9$, $SD = 6.97$). All participants were adults fluent in English who had never been to prison for committing a crime. Due to a noticeable increase in failed attention checks during pilot data collection, the minimal approval rating on Prolific was raised from 90% to 95%. Despite this, data from 46 participants were omitted and replaced due to failed attention checks. Three participants were mistakenly recruited after the intended sample size had been met, and thus, their data were excluded from analyses[1]. Participation took an average of seven minutes.

8.1.2 Materials

Vignettes. All participants responded to the same four vignettes, each detailing a different scenario in which a criminal's eligibility for parole was assessed. Efforts were made to ensure that all four vignettes were similar in length, degree of detail, and contextual aspects, such as how long the criminal had spent in prison and who was described as attributing the traits to the criminal. In each vignette, the criminal's name and pronouns were gender-neutral, their age and ethnicity were not indicated, and their crime and sentence were not specified. The four vignettes can be seen in the Supplementary materials. In the uniquely human socially desirable condition, the criminal was described as *cultured, civilised, sophisticated*, and *moral*, while in the uniquely human condition socially undesirable, the criminal was described as *corrupt, controlling,*

[1]Including data submitted by excess participants in analyses yielded the same results as those reported.

arrogant and *bitter*. In the animalistic desirable condition, the criminal was described as *energetic, trusting, genuine* and *having curiosity*, while in the animalistic undesirable condition, the criminal was described as *uncultured, unrefined, unsophisticated*, and *stupid*.

The following is an example of a vignette describing a criminal with uniquely human, socially desirable traits: "*Alex, known by locals in their hometown as having always been sophisticated, has recently begun their first parole hearing at the local courthouse. Having been tried and convicted 36 months ago, a report by one of the prison's counsellors notes that other prisoners often refer to Alex as being civilised and moral in character. Alex was also described by the counsellor as exhibiting a cultured demeanour since their arrival.*"

Agreement with parole. The dependent variable, agreement with granting parole, was measured using the following single-item measure: '*I think (Alex/Sam/Robin/Jamie) should be granted parole*'. This measure appeared after each vignette, and participants indicated their agreement using an unmarked sliding scale ranging from Strongly Disagree (0) to Strongly Agree (100). The slider's initial starting point was always centred at 50.

Attention check. An additional paragraph describing a criminal named Charlie was included, largely similar to the other four paragraphs. However, in the middle of the paragraph, the following sentence was included: '*This paragraph is an attention check: please move the slider all the way to Strongly Disagree on the left-hand side*'. Data submitted by any participants who did not respond within 20 points of the instructed end of the 100-point scale were omitted and replaced.

8.1.3 Design

This study adopted a 2(trait humanness: uniquely human, shared) × 2(trait desirability: desirable, undesirable) within-subjects factorial design. Counterbalancing ensured that each vignette was associated with each trait category an equal number of times across the participant sample, resulting in four trait-type orders. The trait words were randomly allocated to the position in which they appeared in each vignette using a random order function in Excel. Mirror versions of the trait orders were then created. These two trait-order conditions were also counterbalanced between participants, which was done to control for possible primacy and recency effects of the order in which traits appeared.

8.1.4 Procedure

After participants provided informed consent, they responded to the same demographic questions and inclusion checks as in the other studies. Participants were then shown the first of the four vignettes. After reading the vignette, participants were asked to respond to a single item measuring their agreement with granting parole to the criminal depicted. Participants then repeated the above steps for the remaining three vignettes. The order in which the vignettes were presented to participants was randomised. Each vignette appeared on the screen for at least 15 seconds to maximise the chance that participants read all the relevant information. Participants were debriefed and redirected to Prolific to collect their reward after completing the questionnaire.

8.2 Results

A 2 × 2 within-subjects ANOVA was conducted to examine how variations in the desirability (desirable or undesirable) and humanness (uniquely human or shared with other animals) of the traits used to describe criminals influenced participants' agreement with granting them parole. In line with our prediction, a significant main effect of trait desirability was found, $F(1, 135) = 369.43$, $p = <0.001$, $\eta_p^2 = .732$. Criminals described with socially undesirable traits ($M = 38.82$, $SE = 1.72$) were less likely to be granted parole than were those described with desirable traits ($M = 77.8$, $SE = 1.44$); see Figure 3. A main effect of trait humanness was also found, $F(1, 135) = 51.62$, $p = <0.001$, $\eta_p^2 = .277$. Contrary to the predictions of the dual model, however, criminals who were described with uniquely human traits ($M = 53$, $SE= 1.36$) were less likely to be granted parole than those described with traits shared with other animals ($M = 63.7$, $SE = 1.5$). A significant interaction effect between trait humanness and trait desirability was also found, $F(1, 135) = 54.67$, $p = <0.001$, $\eta_p^2 = .288$. Paired samples t-tests were conducted to examine this interaction effect.

Criminals described using undesirable uniquely human traits (M = 27.9, SE = 1.93) were less likely to be granted parole than were criminals described using desirable uniquely human traits (M = 78, SE = 1.7), $t(135)$ = -20.75, p = <0.001, Cohen's d = -1.78. Similarly, criminals described using undesirable traits shared with other species (M = 49.8, SE= 2.32) were less likely to be granted parole than were those described using desirable traits shared with other species (M = 77.6, SE = 1.6), $t(135)$ = -10.57, p = <0.001, Cohen's d = -0.91.

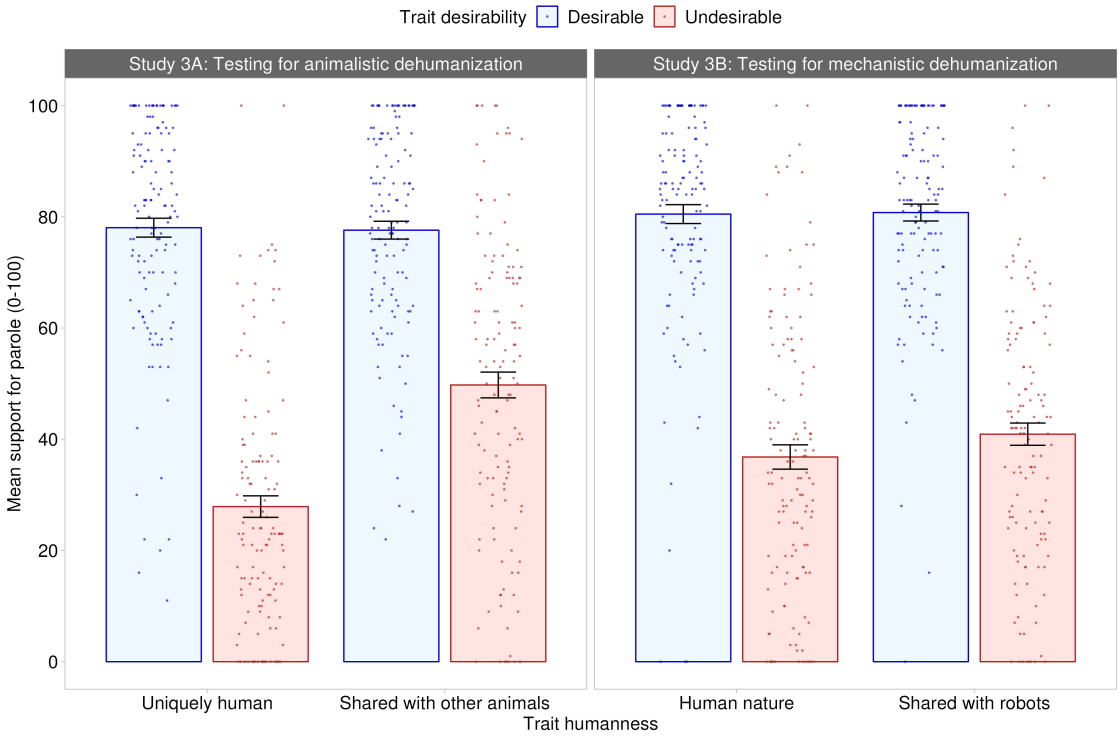

Figure 3: Results of Studies 3A and 3B. Criminals described with undesirable traits were less likely to be granted parole than were those described with desirable traits, regardless of whether or not those traits were uniquely human.

9 Study 3B

Study 3B had an extremely similar design and method to Study 3A. We again employed an experimental manipulation in which we manipulated the perceived humanness and sociality of the traits with which criminals were described and measured how these descriptions influenced participants' parole decisions. In study 3B, we specifically tested for an influence of mechanistic dehumanization by including human nature traits and traits shared with robots in our measures.

As in Study 3A, we predicted that criminals described with undesirable traits would be less likely to be granted parole than those described using desirable traits.

9.1 Methods

9.1.1 Participants

The power analysis described in Study 3A also informed the sample size for Study 3B. A new sample of 136 participants was collected, of whom 76 identified as female, 56 as male, 3 as non-binary, and 1 did not indicate their gender identity. Ages ranged from 18 to 57 (M = 26.4, SD = 8.32). The inclusion criteria were identical to those used in Study 3A, including a minimum Prolific approval rating of 95%. Data from

35 participants were omitted and replaced due to failed attention checks. Five participants were mistakenly recruited after the intended sample size had been met, and thus, their data were excluded from analyses. Participation took an average of just under 7 minutes.

9.1.2 Materials

The agreement with granting parole scale and attention check were the same as those used in Study 3A.

Vignettes. All participants responded to the same four vignettes used in Study 3A but with somewhat different trait words. The desirable human words were *generous, openminded, warm*, and *kind*. The undesirable human words were *jealous, selfish, spiteful*, and *stingy*. The desirable traits shared with robots were *helpful, disciplined, calm* and *efficient*. The undesirable traits shared with robots were *cold, inflexible, superficial*, and *passive*. The following vignette is an example from the undesirable shared condition: "*Sam is currently applying for parole after being convicted of a crime just over three years ago. In assessing Sam's suitability, the parole committee gathered reports from prison staff and other inmates. Guards patrolling the prison grounds noted Sam as being passive. Other prisoners mention Sam as exhibiting superficial behaviour with them for the most part. The prisoner who shares a cell with Sam has referred to them as the most inflexible cell-mate they have ever had. In last week's parole hearing, Sam's responses indicated a cold character*."

9.1.3 Design and procedure

The design and procedure were identical to that of study 3A.

9.2 Results

A 2 × 2 within-subjects ANOVA was conducted to examine how variations in trait humanness (human or shared with robots) and trait desirability (desirable or undesirable) influenced participants' parole decisions. As illustrated in Figure 3, a significant main effect of trait desirability was found, $F(1, 135) = 409.59$, $p = <0.001$, $\eta_p^2 = .752$. Criminals described using undesirable traits ($M = 38.9$, $SE = 1.82$) were less likely to be granted parole than were criminals described using desirable traits ($M = 80.6$, $SE = 1.44$).

No significant main effect of trait humanness was found, $F(1, 135) = 2.78$, $p = .098$, $\eta_p^2 = .020$. Participants were no more likely to grant parole to criminals who were described using human traits ($M = 58.6$, $SE = 1.53$) than those described using traits shared with robots ($M = 60.8$, $SE = 1.33$). Unlike in Study 3A, no interaction effect between trait humanness and trait desirability was found, $F(1, 135) = 2.49$, $p = .117$, $\eta_p^2 = .018$.

10 General Discussion

Across six highly powered and preregistered studies, we examined the hypothesised causal relationship between trait-based dehumanization and harm. The dual model of dehumanization (Haslam, 2006) posits that individuals and groups are sometimes subtly dehumanized by being denied human character traits. To the extent that groups are dehumanized in this way, they are thought to be vulnerable to harm (see Haslam, 2019; Haslam & Loughnan, 2016). The work of Bastian and colleagues (2013) is often cited in support of this claim. Bastian et al. (2013) reported that the fewer human traits participants attributed to criminals, the harsher the punishments participants endorsed for them.

We initially sought to replicate the dehumanization effect reported by Bastian and colleagues (2013) in a conceptually similar design. In Study 1A, we examined the relationship between animalistic and mechanistic dehumanization, as operationalised by Bastian and colleagues, and the harshness of punishment endorsed by participants. In both studies, we successfully replicated previous findings. This demonstrates that our paradigm is sensitive to finding predictive relationships between trait-based dehumanization and harm should they occur.

In Studies 2A and 2B, we investigated the extent to which the previously reported relationship between trait-based dehumanization and harm can be explained by the social desirability of the traits incorporated

into the stimulus set. The dual model (Haslam, 2006) has previously been critiqued for failing to clearly distinguish evidence for trait-based dehumanization from evidence of negative evaluation (Bloom, 2022; Enock, Tipper, et al., 2021; Over, 2021b, 2021a). Bastian and colleagues (2013) operationalised animalistic dehumanization as a reduction in the extent to which participants viewed criminals as possessing traits such as sophistication and refinement. They operationalised mechanistic dehumanization as a reduction in the extent to which participants viewed criminals as possessing traits like warmth and depth. As each of these human traits is socially desirable, it is impossible to determine whether harm is predicted by dehumanization or negative evaluation. In order to tease apart the influence of dehumanization and negative evaluation in harm, we incorporated undesirable human traits into our stimulus set, for example, bitter and spiteful. If trait-based dehumanization explains harm, the previously reported relationship between dehumanization and punishment should remain even when undesirable human traits are incorporated into the stimulus set. If the previously reported relationship is better explained by negative evaluation, then trait desirability should moderate the relationship with punishment. In support of the latter claim, Studies 2A and 2B showed that the more desirable human traits participants attributed to criminals, the less harshly participants thought they should be punished. The more undesirable human traits participants attributed to criminals, the more harshly participants thought they should be punished.

In Study 3, we sought to further distinguish between these two competing hypotheses using an experimental manipulation. In Studies 3A and 3B, we described criminals in traits that varied in perceived humanness and sociality and measured the influence of these varying descriptions on participants' parole decisions. This experimental design allowed us to directly test the hypothesised causal relationship between trait-based dehumanization and punishment. Converging with the findings of Study 2, we found that criminals described with undesirable traits were less likely to be granted parole than were criminals described with desirable traits, regardless of whether or not those traits were uniquely human. There was no evidence for the hypothesis that criminals described with uniquely human terms would be more likely to be granted parole.

These findings fit with broader critiques of social psychological models of dehumanization. Enock and colleagues (2021) showed that what appears to be evidence for trait-based dehumanization of immigrants and political groups is better explained by negative evaluation. Similarly, Enock, Flavell, et al. (2021) presented evidence that what appears to be emotion-based dehumanization of seven different outgroups is better explained by negative evaluation. In these studies, participants were more likely to attribute prosocial emotions to the ingroup regardless of whether they were uniquely human or not. Participants were more likely to attribute antisocial emotions to the outgroup, regardless of whether they were uniquely human or not. In further work, Enock & Over (2022) presented evidence that the apparent relationship between emotion-based dehumanization and reductions in prosocial behaviour is better explained by negative evaluation.

Partially in response to these critiques, Kteily & Landry (2022) presented a new social psychological model of dehumanization in which to dehumanize an individual or group is to perceive them as less than the ideal human. Under this characterisation of dehumanization, to view a group as possessing negative attributes is to dehumanize them. However, to define dehumanization in such a broad way as any negative evaluation renders almost all social judgments dehumanizing (Bloom, 2022). It seems unlikely that we dehumanize our closest and most loved kin simply by perceiving their imperfections. It is crucial that future conceptual research on dehumanization more clearly delineates dehumanization from negative evaluation (Bloom, 2022; Over, 2021b, 2021a).

It is important to acknowledge that we considered only one target group in this study - criminals. We based this decision on the influence the findings of Bastian et al. (2013) have had on the literature. However, there may be more evidence for the hypothesised causal relationship between trait-based dehumanization and harm in other intergroup contexts. In addition to examining additional intergroup contexts, future research should also incorporate more trait terms into stimulus sets. Research on dehumanization has been critiqued for using relatively small stimulus sets (Vaes, 2023). Indeed, some studies have used a single trait term to assess dehumanization. For example, Leidner et al. (2013) measured dehumanization by asking participants to rate the extent to which they agreed that members of the target outgroup experienced compassion. It will always remain possible that evidence for the causal relationship between dehumanization and harm could be found with a more sensitive paradigm.

We are not trying to argue that trait-based dehumanization never occurs. Rather, our argument is considerably more modest. Taken in conjunction with other recent results, it is apparent that evidence for trait-based dehumanization has often been confounded with evidence for negative evaluation (Bloom, 2022; Enock, Tipper, et al., 2021; Enock, Flavell, et al., 2021; Over, 2021b, 2021a). The results of the current study add to this growing body of critiques by showing that the findings of Bastian et al. (2013), often cited as key evidence for the claim that trait-based dehumanization leads to harm, are considerably less convincing than they first appear. It is imperative that future research tests whether there is evidence for trait-based dehumanization when trait desirability is controlled for, given the grave importance of understanding predictors of intergroup harm in the real world.

Ethical statement. All studies received ethical approval from the Psychology Departmental Ethics Committee at the University of York (approval number 926). Informed consent was obtained from participants at the start of each experimental session according to approved ethical procedures. All studies were performed in accordance with the relevant guidelines and regulations.

Data accessibility. All studies reported in this article were pre-registered and the data is available open access. Data files, pre-registration documents, a fully computationally reproducible version of the manuscript, and supplementary materials, including the stimuli used for each study, can be found at http://doi.org/10.17605/OSF.IO/D4CVP.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. R.A.B.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing—original draft, writing—review and editing; F.E.E.: conceptualization, formal analysis, investigation, methodology, resources, supervision, validation, visualization, writing—review and editing; H.O.: conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing—review and editing. All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research was supported by the European Research Council under the European Union's Horizon 2020 Programme, grant number ERC- STG-755719 awarded to HO.

Acknowledgements. Production of the reproducible version of this manuscript was supported by an Enhancing Research Culture award from Research England.

References

Andrighetto, L., Baldissarri, C., Lattanzio, S., Loughnan, S., & Volpato, C. (2014). Human-itarian aid? Two forms of dehumanization and willingness to help after natural disasters. *British Journal of Social Psychology*, *53*(3), 573–584. https://doi.org/10.1111/BJSO.12066

Bain, P., Vaes, J., Kashima, Y., Haslam, N., & Guan, Y. (2011). Folk conceptions of humanness: Beliefs about distinctive and core human characteristics in australia, italy, and china. *Journal of Cross-Cultural Psychology*, *43*(1), 53–58. https://doi.org/10.1177/0022022111419029

Banton, O., West, K., & Kinney, E. (2020). The surprising politics of anti-immigrant prejudice: How political conservatism moderates the effect of immigrant race and religion on infrahumanization judgements. *British Journal of Social Psychology*, *59*(1), 157–170. https://doi.org/10.1111/BJSO.12337

Barber, M., & Davis, R. (2022). Partisanship and the trolley problem: Partisan willingness to sacrifice members of the other party. *Research & Politics*, *9*(4). https://doi.org/10.1177/20531680221137143

Bastian, B., Denson, T. F., & Haslam, N. (2013). The roles of dehumanization and moral outrage in retributive justice. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061842

Bastian, B., & Haslam, N. (2010). Excluded from humanity: The dehumanizing effects of social ostracism. *Journal of Experimental Social Psychology*, *46*(1), 107–113. https://doi.org/10.1016/j.jesp.2009.06.022

Bloom, P. (2017). The root of all cruelty? Perpetrators of violence, we're told, dehumanize their victims. The truth is worse. *The New Yorker*.

Bloom, P. (2022). If everything is dehumanization, then nothing is. *Trends in Cognitive Sciences*, *26*(7), 539. https://doi.org/10.1016/J.TICS.2022.03.001

Brennan, R. A., Enock, F. E., & Over, H. (n.d.). [dataset].

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*(5), 645–662. https://doi.org/10.1002/EJSP.2357

Chen-Xia, X. J., Betancor, V., Rodríguez-Gómez, L., & Rodríguez-Pérez, A. (2023). Cultural variations in perceptions and reactions to social norm transgressions: A comparative study. *Frontiers in Psychology*, *14*(1243955). https://doi.org/10.3389/fpsyg.2023.1243955

Decker, K. A., & Lord, C. G. (2023). Self-polarization: Lionizing those who agree and demonizing those who disagree. *Basic and Applied Social Psychology*, *45*(5), 125–137. https://doi.org/10.1080/01973533.2023.2234534

Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. M. (2010). The SAGE handbook of prejudice, stereotyping and discrimination. *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, 1–663. https://doi.org/10.4135/9781446200919

Enock, F. E., Flavell, J. C., Tipper, S. P., & Over, H. (2021). No convincing evidence outgroups are denied uniquely human characteristics: Distinguishing intergroup preference from trait-based dehumanization. *Cognition*, *212*, 1–73. https://doi.org/10.1016/j.cognition.2021.104682

Enock, F. E., & Over, H. (2022). Reduced helping intentions are better explained by the attribution of antisocial emotions than by "infrahumanization." *Scientific Reports*, *12*(1), 1–14. https://doi.org/10.1038/s41598-022-10460-0

Enock, F. E., & Over, H. (2023). Animalistic slurs increase harm by changing perceptions of social desirability. *Royal Society Open Science*, *10*(7), 230203. https://doi.org/10.1098/rsos.230203

Enock, F. E., Tipper, S. P., & Over, H. (2021). Intergroup preference, not dehumanization, explains social biases in emotion attribution. *Cognition*, *216*. https://doi.org/10.1016/j.cognition.2021.104865

Esses, V. M., Veenvliet, S., Hodson, G., & Mihic, L. (2008). Justice, morality, and the dehumanization of refugees. *Social Justice Research*, *21*(1), 4–25. https://doi.org/10.1007/s11211-007-0058-4

Farr, R. M. (1996). *The roots of modern social psychology, 1872–1954: Vol. xvii*. Blackwell publishing.

Gaines, S. O., & Reed, E. S. (1995). Prejudice: From allport to DuBois. *American Psychologist*, *50*(2), 96–103. https://doi.org/10.1037/0003-066X.50.2.96

Giner-Sorolla, R., Burgmer, P., & Demir, N. (2021). Commentary on over (2021): Well-taken points about dehumanization, but exaggeration of challenges. *Perspectives on Psychological Science*, *16*(1), 24–27. https://doi.org/10.1177/1745691620953788

Goldenberg, J. L., Courtney, E. P., & Felig, R. N. (2021). Supporting the dehumanization hypothesis, but under what conditions? A commentary on over (2021). *Perspectives on Psychological Science*, *16*(1), 14–21. https://doi.org/10.1177/1745691620917659

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/SCIENCE.1134475

Hare, B., & Woods, V. (2020). *Survival of the friendliest: Understanding our origins and rediscovering our common humanity*. Random House.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological Science*, *17*(10), 847–853. https://doi.org/10.1111/J.1467-9280.2006.01793.X

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Haslam, N. (2019). The many roles of dehumanization in genocide. *Confronting Humanity at Its Worst*, 119–138. https://doi.org/10.1093/OSO/9780190685942.003.0005

Haslam, N. (2021). The social psychology of dehumanization. In M. K (Ed.), *The routledge handbook of dehumanization*. https://doi.org/10.4324/9780429492464-chapter8

Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, *30*(12), 1661–1673. https://doi.org/10.1177/0146167204271182

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology*, *65*, 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

Haslam, N., & Loughnan, S. (2016). How dehumanization promotes harm. In G. Miller Arthur (Ed.), *The*

social psychology of good and evil (pp. 140–158). Guilford Publications.
Kasper, A., Frébert, N., & Testé, B. (2022). Caught COVID-19? covidiot! *Social Psychology*, *53*(2), 84–95. https://doi.org/10.1027/1864-9335/a000478
Kteily, N. S., & Landry, A. P. (2022). Defining dehumanization broadly does not mean including everything. *Trends in Cognitive Sciences*, *26*(7), 540–541. https://doi.org/10.1016/J.TICS.2022.04.003
Lang, J. (2010). Questioning dehumanization: Intersubjective dimensions of violence in the nazi concentration and death camps. *Holocaust and Genocide Studies*, *24*(2), 225–246. https://doi.org/10.1093/HGS/DCQ026
Lang, J. (2020). The limited importance of dehumanization in collective violence. *Current Opinion in Psychology*, *35*, 17–20.
Leidner, B., Castano, E., & Ginges, J. (2013). Dehumanization, retributive and restorative justice, and aggressive versus diplomatic intergroup conflict resolution strategies. *Personality and Social Psychology Bulletin*, *39*(2), 181–192. https://doi.org/10.1177/0146167212472208
Leyens, J. P., Cortes, B., Demoulin, S., Dovidio, J. F., Fiske, S. T., Gaunt, R., Paladino, M. P., Rodriguez-Perez, A., Rodriguez-Torres, R., & Vaes, J. (2003). Emotional prejudice, essentialism, and nationalism: The 2002 Tajfel Lecture. *European Journal of Social Psychology*, *33*(6), 703–717. https://doi.org/10.1002/EJSP.170
Leyens, J. P., Paladino, P. M., 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*(2), 186–197. https://doi.org/10.1207/S15327957PSPR0402_06
Leyens, J. P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M. P., Vaes, J., & Demoulin, S. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology*, *31*(4), 395–411. https://doi.org/10.1002/EJSP.50
Loughnan, S., Haslam, N., Sutton, R. M., & Spencer, B. (2013). Dehumanization and social class. *Social Psychology*, *45*(1), 54–61. https://doi.org/10.1027/1864-9335/A000159
Major, B., & Sawyer, P. J. (2009). Attributions to discrimination: Antecedents and consequences. In *Handbook of prejudice, stereotyping, and discrimination* (pp. 89–110).
Manne, K. (2016). Humanism: A critique. *Social Theory and Practice*, *42*(2), 389–415. https://doi.org/10.5840/SOCTHEORPRACT201642221
Manne, K. (2018). Down girl: The logic of misogyny. In *Down Girl: The Logic of Misogyny*. https://doi.org/10.1093/oso/9780190604981.001.0001
Morehouse, K. N., Maddox, K., & Banaji, M. R. (2023). All human social groups are human, but some are more human than others: A comprehensive investigation of the implicit association of "human" to US racial/ethnic groups. *Psychological and Cognitive Sciences*, *120*(22). https://doi.org/10.1073/pnas.230099512
Over, H. (2021a). Falsifying the dehumanization hypothesis. *Perspectives on Psychological Science*, *16*(1), 33–38. https://doi.org/10.1177/1745691620969657
Over, H. (2021b). Seven challenges for the dehumanization hypothesis. *Perspectives on Psychological Science*, *16*(1), 3–13. https://doi.org/https://doi.org/10.1177%2F1745691620902133
Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(32), 8511–8516. https://doi.org/10.1073/pnas.1705238114
Rousseau, D. L., Gorman, B., & Baranik, L. E. (2023). Crossing the line: Disgust, dehumanization, and human rights violations. *Socius*, *9*, 23780231231157686. https://doi.org/10.1177/23780231231157686
Ruiter, de A. (2022). Failing to see what matters most: Towards a better understanding of dehumanisation. *Contemporary Political Theory 2022*, 1–22. https://doi.org/10.1057/S41296-022-00569-2
Smith, D. L. (2011). *Less than human: Why we demean, enslave, and exterminate others* (pp. 1–336). St. Martin's Griffin.
Smith, D. L. (2016). Paradoxes of dehumanization. *Social Theory and Practice*, *42*(2), 416–443. https://doi.org/10.5840/SOCTHEORPRACT201642222
Smith, D. L. (2020). *On Inhumanity: Dehumanization and How to Resist It*. Oxford University Press. https://doi.org/10.1093/oso/9780190923006.001.0001

Smith, D. L. (2021). *Making Monsters: The Uncanny Power of Dehumanization*. Harvard University Press. https://doi.org/10.4159/9780674269781

Vaes, J. (2023). Dehumanization after all: Distinguishing intergroup evalutation from trait-based dehumanization. *Cognition*, *231*, 105329. https://doi.org/10.1016/J.COGNITION.2022.105329

Vaes, J., Paladino, M. P., & Haslam, N. (2021). Seven clarifications on the psychology of dehumanization. *Perspectives on Psychological Science*, *16*(1), 28–32. https://doi.org/10.1177/1745691620953767

West, S. J., & Thomson, N. D. (2022). Identifying the emotions behind apologies for severe transgressions. *Motivation and Emotion*, *47*, 257–269. https://doi.org/10.1007/s11031-022-09993-8