Running head: THE AGENCY OF HARMFUL AGENTS

Perceiving the agency of harmful agents:

A test of dehumanization versus moral typecasting accounts

Forthcoming (2016) in the *Cognition*

Mansur Khamitov

Ivey Business School, Western University, Canada

Jeff D. Rotman*

Ivey Business School, Western University, Canada

Jared Piazza

Department of Psychology, Lancaster University, United Kingdom

WORD COUNT: 14,879

Author Note:

The first two authors contributed equally to this paper, and order of authorship was determined randomly. We thank Paul Conway and James Olson for their helpful comments on an earlier version of this paper.

* Correspondence concerning this article should be addressed to:

Jeff D. Rotman, Ivey Business School, Western University, London, ON, N6G 0N1, Canada. Tel.: +1 (519) 670 4985. E-mail address: jrotman.phd@ivey.ca

Abstract

It is clear that harmful agents are targets of severe condemnation, but it is much less clear how perceivers conceptualize the agency of harmful agents. The current studies tested two competing predictions made by moral typecasting theory and the dehumanization literature. Across six studies, harmful agents were perceived to possess *less agency* than neutral (non-offending) and benevolent agents, consistent with a dehumanization perspective but inconsistent with the assumptions of moral typecasting theory. This was observed for human targets (Studies 1-2b, and 4-5) and corporations (Study 3), and across various gradations of harmfulness (Studies 3-4). Importantly, denial of agency to harmful agents occurred even when controlling for perceptions of the agent's likeability (Studies 2a and 2b) and while using two different operationalizations of agency (Study 2a). Study 5 showed that harmful agents are denied agency primarily through an inferential process, and less through motivations to see the agent punished. Across all six studies, harmful agents were deemed less worthy of moral standing as a consequence of their harmful conduct and this reduction in moral standing was mediated through reductions in agency. Our findings clarify a current tension in the moral cognition literature, which have direct implications for the moral typecasting framework.

Keywords: Harmfulness, Agency, Moral Standing, Dehumanization, Moral Typecasting, Moral Cognition

Perceiving the agency of harmful agents:

A test of dehumanization versus moral typecasting accounts

1. Introduction

There is probably no moral intuition more fundamental and ubiquitous than the rejection of cruelty or the infliction of harm for purely selfish reasons (Gert, 2004; Graham, Haidt, & Nosek, 2009; Greene, 2012; Henrich et al., 2006; Piazza, Landy, & Goodwin, 2014; Pinker, 2012; Sinnott-Armstrong, 2009; Sousa, Holbrook, & Piazza, 2009; Sousa & Piazza, 2014; Turiel, 1983). Historically, societies have not always agreed on which actions constitute cruelty or which individuals and entities are deserving of protection from such abuses (Haslam & Loughnan, 2014; Piazza et al., 2014; Singer, 2011). Yet this fact does not negate the core intuition that individuals who cause unjustified harm have violated an implicit social contract to respect the basic interests of others (Baumard, Andre, & Sperber, 2013; Sousa & Piazza, 2014), or the retributive logic that harmful agents are deserving of punishment (Ashworth, 2010; Baumard, 2011; Carlsmith, Darley, & Robinson, 2002; Darley & Pittman, 2003).

A vast literature within psychology supports the idea that harmful agents are targets of often severe condemnation (e.g., Bastian, Denson, & Haslam, 2013; Bastian, Laham, Wilson, Haslam, & Koval, 2011; Carlsmith et al., 2002; Gray, 2014; Gray & Wegner, 2009; Vasquez, Loughnan, Gootjes-Dreesbach, & Weger, 2014). However, much less research has considered the attributions people make with regards to the underlying agency of harmful agents. Currently, there are two perspectives on the matter, each with competing predictions. According to moral typecasting theory (hereon MTT; Gray & Wegner, 2009), harmful agents should be perceived as highly agentive - indeed, *as agentive* as positive moral actors - and certainly *more agentive* than neutral or non-offending actors (see also Gray, 2010; Gray & Schein, 2012; Gray & Wegner,

2011). From this perspective, when a person commits an act of cruelty they are "transformed" (Gray, 2010) or "typecasted" (Gray & Wegner, 2009) in the eyes of those bearing witness to their actions. The result is that perceivers imbue the target with agency (see Gray, 2010), or, put another way, they are attributed the qualities befitting a "moral agent" (see Gray & Wegner, 2009). Such qualities might include, "self-control, morality, memory, emotion recognition, planning, communication, and thought" (Gray & Wegner, 2009, p. 506; see also Gray, Gray, & Wegner, 2007).

From a different perspective, however, harmful agents should not be typecasted as agents, but *denied agency*, as an extension of the human inclination to dehumanize cruel agents (Bastian et al., 2013; Haslam & Loughnan, 2014; Leidner, Castano, & Ginges, 2013; Viki, Fullerton, Raggett, Tait, & Wiltshire, 2012). According to work on dehumanization, harmful agents are often seen by others as lacking basic aspects of humanity or "humanness" (Haslam, 2006; Haslam, Kashima, Loughnan, Shi, & Suitner, 2008), such as civility and warmth, and at times may even be imbued with animalistic or machine-like traits (Bastian et al., 2013; Vasquez et al., 2014). Thus, currently, there exists a tension in the psychological literature regarding how harmful agents are conceptualized. In the present set of studies, we show that the moral typecasting hypothesis that harmful agents are typecasted as agentive fails to hold up to empirical scrutiny. Rather than being typecasted as moral agents, we show that harmful agents are *denied* agency, along with other aspects of their humanity.

1.1 Moral typecasting, dehumanization, and defining agency

MTT puts forth the provocative claim that agents that inflict harm on others, and, likewise, agents who do good deeds for others, are typecasted as "moral agents" and not "moral patients," i.e., they are ascribed the qualities befitting an agent, such as rationality and self-

control, but not the qualities befitting a patient or victim, such as the capacity to suffer (Gray & Wegner, 2009). Conversely, according to MTT, individuals who are victimized are typecasted as moral patients, but not as moral agents, and are thus ascribed the qualities befitting a patient, but not the qualities befitting an agent. In the present paper we focus empirically on the former half of the claim: the typecasting of moral agents as agentive.

One of the difficulties with interpreting the moral typecasting hypothesis involves the various ways in which agency has been operationally defined (for a thorough review, see Piazza et al., 2014). In the literature on mind perception, agency is often defined in terms of "higher" cognitive capacities, such as being able to reason, communicate, exert self-control, imagine, and plan one's actions (see especially Gray et al., 2007; but also Gray & Schein, 2012; Gray, Waytz, & Young, 2012; Gray & Wegner, 2012; Haslam et al., 2008; Waytz, Gray, Epley, & Wegner, 2010). Consistent with the mind perception literature, Sytsma and Machery (2012) operationalized agency in terms of higher intelligence, which includes such traits as language, creativity, and the capacity for sophisticated culture (e.g., music, poetry). Another perspective from social psychology defines agency more broadly in terms of being active, tenacious, effective at pursing one's goals, and having control over one's environment (Abele, Uchronski, Suitner, & Wojciszke, 2008; Abele & Wojciszke, 2007). Indeed, Gray and Wegner (2009) also suggest there are "general" aspects of agency (e.g., being "determined, "powerful") that might be ubiquitous to all agents (see Gray & Wegner, 2009, Study 4b). Thus, there are several perspectives on agency and its definition, with research revealing at least two important aspects: intelligence (or "cognition" broadly defined) and the capacity for effective goal-directed activity (see Piazza et al., 2014).

If we turn to the manner in which researchers from MTT have defined agency, we find a certain degree of inconsistency in the way agency is defined and operationalized. Gray and Wegner (2009) are quite clear that they see the moral typecasting hypothesis as compatible with the definition of agency coming from the mind perception literature (see Footnote 1 for one illustrative quotation). On the other hand, in their studies Gray and Wegner (2009) assessed moral agency using quite a limited set of measures pertaining to intentional action (intentionality) and blame and praise (culpability), as opposed to the broader, richer conception of agency identified by the mind perception literature (see also Gray & Wegner, 2011). Intentionality is only one aspect of agency among many, and, arguably, blame/praise has more to do with the potential consequences of perceiving agency rather than the direct possession of agency. Nevertheless, it has been concluded on the basis of these limited measures that harmful agents (and benevolent agents) are perceived as agentive (see Footnote 2 for illustrative quotations). Furthermore, because Gray and Wegner did not assess agency in a comprehensive manner it is not at all clear whether perceptions of the actors' agency within these studies are truly responsible for the attribution of blame and intentionality. Some recent research suggests attributions of intentionality and blame are, at times, separable from the activity (or inactivity) that brought about the harmful outcome (e.g., see Critcher, Inbar, & Pizarro, 2013; Cushman, Knobe, & Sinnott-Armstrong, 2008; Knobe, 2003). Intentionality merely requires the perception that an act is goal directed (i.e., desired and intended; Malle & Knobe, 1997); it does not require the attribution of high levels of agency—for example, high levels of rationality, imagination, or self-control. Thus, perceivers may at times perceive intentionality despite a deficit of agency on the part of the agent, such as when an agent's thoughtless actions have unintended, harmful consequences (Knobe, 2003). Likewise, neither do attributions of blame require high levels of

agency (Malle, Guglielmo, & Monroe, 2014); a person can be held morally blameworthy for a misdeed without exerting much agency (i.e., planning, rationality, self-control, etc.), such as when someone causes harm impulsively (Critcher et al., 2013), without a good reason (Darley, Klosson, & Zanna, 1978), or as a side effect of another action (Knobe, 2003). In such cases attributions of blame may arise simply as a matter of procedural justice for causing foreseeable harm or because the agent's lack of agency (e.g., lack of rationality or self-control) suggests a deficiency in the agent's character which poses an ongoing threat to others (Critcher et al., 2013). It should be noted, however, that low agency may at times serve to *mitigate* blame as well—for example, when the agent is mentally impaired (Christopher & Pinals, 2010; Hart, 1968). Given the complex relationship between agency, intentionality, and blame, the use of intentionality and culpability as the methodological standard for testing the moral typecasting hypothesis is somewhat problematic. A richer and more direct test of the moral typecasting hypothesis would be to assess agency traits more comprehensively in terms of the capacity for rationality, planning, self-control, imagination, emotion recognition, and so on, after manipulating perceptions of the agent's harmfulness.

When issues pertaining to the assessment of agency are addressed we can see an empirical tension emerging for the moral typecasting hypothesis. The notion that harmful agents are typecasted as moral agents appears at odds with the psychological literature on dehumanization (e.g., Bastian et al., 2013; Bastian et al., 2011; Haslam et al., 2008; Haslam & Loughnan, 2014; Viki et al., 2012). This literature suggests that agents who cause others harm (e.g., criminals, sex offenders) are often denied important human characteristics, including qualities related to agency (e.g., rationality) and patiency (e.g., being emotionally responsive). For example, in one study (Bastian et al., 2013), participants read about various types of criminal

acts ranging in severity (white-collar crime, violence, child molestation). Then they ascribed the actors who committed these crimes various traits largely pertaining to agency and patiency - eight traits derived from Bastian and Haslam (2010) having to do with the rationality, refinement, and civility of the target ("human uniqueness" traits), or having to do with the emotional depth, warmth, and responsiveness of the target ("human nature" traits). The authors found that the ascription of human uniqueness and human nature traits decreased proportionate to the perceived severity of the offender's crime. Thus, those who committed more severe crimes were denied traits, arguably constitutive of agency and patiency, relative to those who committed less severe crimes.

Why might harmful agents be denied agency? There are at least two plausible mechanisms. First, perceivers might *infer* from the behavior of harmful agents that they lack agentive traits, such as rationality, intelligence, or self-control. In other words, the denial of agentive traits may be an inference perceivers derive from the harmful behavior displayed by the agent. If this is the case, then we should be able to strengthen the denial of agency under conditions where we manipulate the rationality of an agent's motivations for causing harm. Specifically, we would expect perceivers to attribute *less agency* to targets who have less rational motivations for causing harm (e.g., causing harm for the fun of it) compared to agents with more rational motivations (e.g., causing harm for more utilitarian reasons).

A second plausible mechanism is moral disengagement (see Bandura, 1999; Castano & Giner-Sorolla, 2006; Haslam & Loughnan, 2014; Leidner et al., 2013; Leidner, Castano, Zaiser, & Giner-Sorolla, 2010). Perceivers may deny agency to harmful agents as a means of legitimating any aggressive action taken against the agent as retribution for their crime. From this perspective, the denial of agency would aid in preempting any guilt one might experience

when retaliating against harmful offenders. Dehumanization in the context of moral disengagement has generally been understood as a response to interpersonal or intergroup *victimization*—for example, members of a group may deny the full humanity of a victimized group as a means of reducing collective guilt and preserving positive views about the morality of one's own group (Castano & Giner-Sorolla, 2006; Leidner et al., 2010). Although less work has looked at moral disengagement in the context of punishing offenders, it is possible that a motivated use of dehumanization might also occur in this context. Since punishment often takes the form of aggressive action, incapacitation, or the deprivation of basic rights and freedoms, denying offenders full humanity may help legitimate taking action against them. Indeed, the study of Bastian et al. (2013), reviewed above, is consistent with a moral disengagement perspective (though it is also consistent with an inferential account): the harmful agents (e.g., violent offenders) in their study were dehumanized and it was the degree to which the targets were dehumanized that predicted the severity of retributive justice directed at them.

If the moral disengagement hypothesis is correct, then denial of agency to harmful agents should be amplified when motivations to see the agent punished are made salient. It follows that if people dehumanize harmful agents as a means of legitimating punitive reprisals we would expect people to increase their level of dehumanization—including denying them high levels of agency—as a function of their desire to see the perpetrator brought to justice.

1.2 The present studies and hypotheses

The current research presents six studies that sought to test the competing predictions made by moral typecasting theory and the dehumanization literature regarding the manner in which lay individuals conceptualize harmful agents, particularly with regards to perceptions of their agency. According to MTT, agents who cause others harm should be attributed greater

levels of agency than neutral agents (i.e., non-offenders), on par with levels of agency attributed to benevolent agents, as this is part of what it means to be typecast as a moral agent. For example, Gray (2010, p. 253) writes, "…those who do moral deeds, whether laudable or heinous, are perceived as relatively higher in agency and lower in experience (…) those who help or harm others not only are perceived to be more agentic but also are permanently 'typecast' as such" (also see Footnote 3). By contrast, from a dehumanization perspective, harmful agents should be attributed *less* agency than *both* neutral agents and benevolent agents. Across six studies, we contrasted these competing accounts using human targets (Studies 1-2b, and 4-5) as well as a non-human target (a corporation; Study 3). We operationalised agency both in terms of mind attributions (Gray et al., 2007; Studies 1-5) and traits related to activity level/tenacity (Piazza et al., 2014; Study 2a). In Study 2 we ruled out likeability, both statistically (Study 2a) and experimentally (Study 2b) as an alternative explanation for why harmful agents are denied agency, and, in Studies 3 and 4, we examined agency attributions across several gradations of harm. We also investigated, in Study 5, the potential mechanisms (the inferential and motivational mechanisms discussed above) involved in the denial of agency to harmful agents.

As a secondary goal, we sought to determine whether agency denial contributes to judgments of the agent's moral standing (e.g., whether the agent should have its interests protected). Past research by Piazza et al. (2014) have found that harmful agents are ascribed less moral standing than neutral and benevolent agents. These authors showed that it is the perceived *harmful character* of the target, and not the perceived agency, that influences judgments of moral standing at least among non-human animal targets. However, in those studies, agency was narrowly defined in terms of the agent's capacity to effectively act upon and bring about one's goals. Furthermore, the focus was on non-human animals, thus, it remains to be seen whether

perceptions of agency, when broadly defined in terms of various mental capacities, might mediate the relationship between harmfulness and the moral standing of *human* targets. We might speculate that perceptions of agency play a larger role in the moral evaluations of humans, since human beings, as a species, are generally perceived to possess higher base rates of agency (see Gray et al., 2007), and thus any downward adjustment of agency will be easier to detect.

2. Study 1: The Agency of Harmful People

Study 1 provided an initial test of the competing hypotheses of moral typecasting theory and dehumanization theory using human individuals as targets. We manipulated the perceived harmfulness of the agent in a between-subjects design (harmful vs. neutral vs. benevolent) and had participants rate the target on agency traits, taken from the mind perception literature, and form judgments of the target's moral standing. We predicted, consistent with a dehumanization perspective, that harmfulness would reduce judgments of an agent's moral standing, and that this reduction in moral standing would be mediated by attributions of agency.

2.1 Method

2.1.1 Participants

Participants were 60 American adults (42 male; 18 female; M_{age} = 30.41 years, SD = 10.54) recruited via Amazon's Mechanical Turk. Participants were compensated $.25 and the average completion time was two minutes and thirty-two seconds. Recruitment was limited to people located in the United States.

2.1.2 Materials and procedures

Study 1 manipulated agent category via a vignette methodology in a between-subjects design. Participants were randomly assigned to one of the three agent conditions (harmful; neutral; benevolent). In all three conditions, participants first read some filler information about

the target: "John is a 23 year old male. He is about six feet tall, with brown hair and blue eyes." Then they read additional information, which varied depending on the agent condition they were assigned to:

> *Harmful agent*: John has taken to crime. He spends his nights mugging individuals, often stealing everything they have and leaving them injured. He also has a girlfriend, but cheats on her regularly.
>
> *Benevolent agent*: John has taken to giving to charity. He spends his nights providing food and shelter to individuals. He also has a girlfriend and is very loyal to her.

Participants in the *neutral agent* condition were told that John was currently working in retail. Following the manipulation, participants answered a manipulation check on the perceived harmfulness of the agent—five adjectives ($\alpha = .97$; see Appendix A for a full list of items used across studies)—adapted from Piazza et al. (2014, Study 2). Each adjective was rated on a 7-point scale in terms of ''the extent to which the agent has the following qualities'' (*Not at all* to *Completely*). Perceptions of the target's agency was assessed using seven cognitive trait items taken from Gray et al. (2007), for example, "capable of making plans and working towards goals" ($\alpha = .93$). Agreement with these items was rated on a 1 (*Strongly disagree*) to 7 (*Strongly agree*) scale. Lastly, participants judged the moral standing of the agent using Piazza et al.'s (2014, Study 2) 5-item moral standing scale, adapted for use with human targets (e.g., "To what extent do you think this person deserves to be protected from harm?" $\alpha = .94$; all on a 1–7 scale, *Not at all* to *Extremely*). For one of the moral standing items "kill" was replaced with "steal" to soften it. No other measures were collected aside from basic demographics.

2.2 Results

2.2.1 Manipulation check

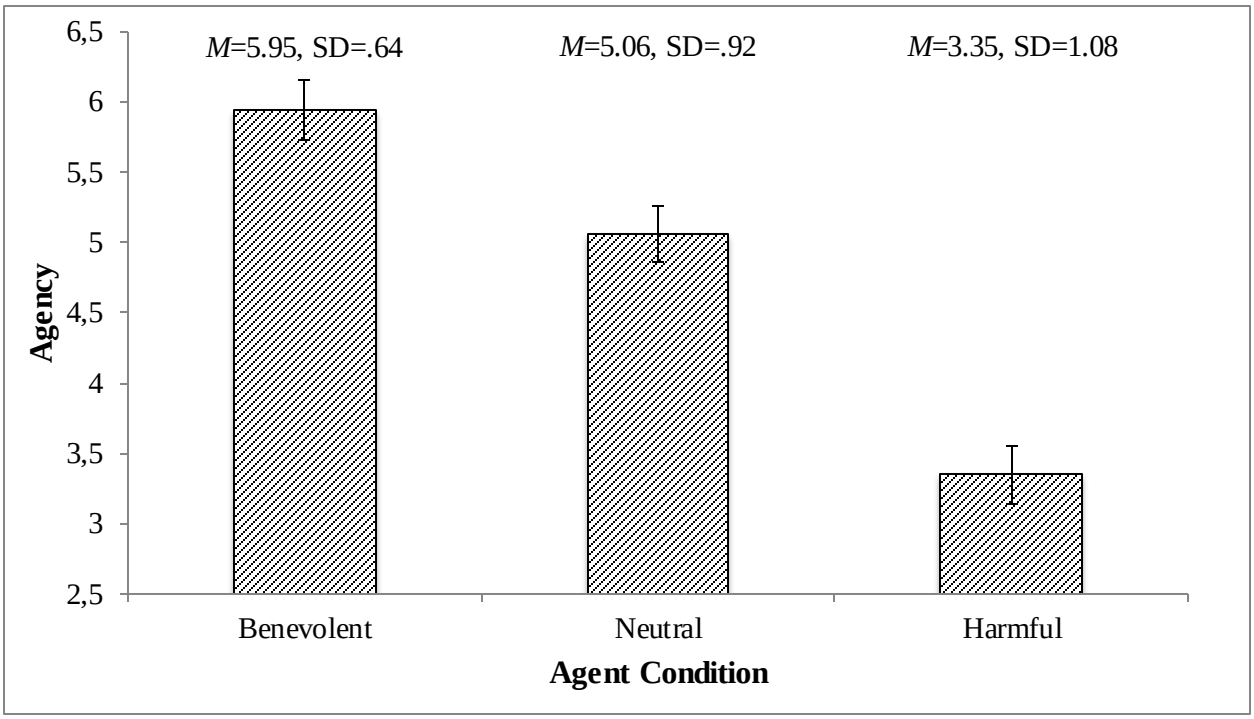
The harmfulness manipulation produced a large effect in the expected direction, $F(2,56) = 91.25$, $p < .001$, $\eta^2_p = .765$. The target was rated as most harmful in the harmful agent condition ($M = 5.99$, SD = 1.08) and least harmful in the benevolent agent condition ($M = 1.88$, SD = 1.01), with the target in the neutral agent condition falling in the middle ($M = 3.33$, SD = .76). All pairwise comparisons were significantly different at $p < .001$.

2.2.2 Main analysis and mediation

Two one-way ANOVAs were conducted on the agency index and, separately, moral standing index. There was a large main effect of agent condition on the agency index, $F(1,57) = 39.45$, $p < .001$, $\eta^2_p = .594$. Consistent with a dehumanization effect, John was attributed the lowest levels of agency in the harmful agent condition and the highest levels of agency in the benevolent agent condition, with the neutral agent condition falling in the middle (see Figure 1 for means and standard deviations [SDs]). All pairwise comparisons were significantly different at $p < .003$. There was a significant main effect of agent condition on moral standing ratings, $F(1,55) = 19.86$, $p < .001$, $\eta^2_p = .428$. John was attributed lower moral standing in the harmful agent condition ($M = 4.02$, SD = 1.35) than in the benevolent ($M = 6.12$, SD = 1.04), $p < .001$, and neutral agent conditions ($M = 5.89$, SD = .91), $p < .001$. John was attributed similar levels of moral standing in the benevolent and neutral agent conditions, $p > .54$.

Figure 1

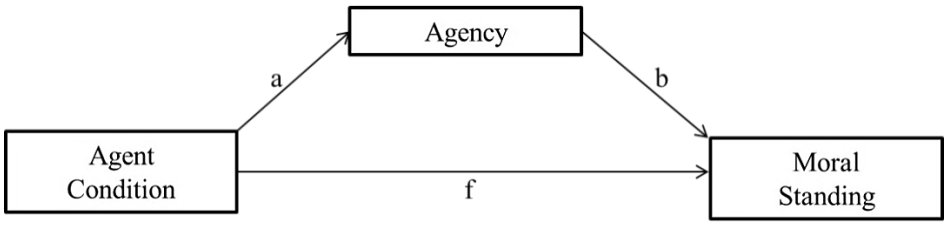
Attributions of agency to the human target by agent condition (Study 1). Error bars ± 1 SE.



We used the PROCESS macro to conduct a mediation analysis (Hayes, 2013), with a bootstrapping procedure (5,000 resamples) to construct bias-corrected confidence intervals. The bootstrapping procedure showed that, as predicted, the influence of the agent condition (1 = benevolent; 2 = neutral; 3 = harmful) on moral standing scores was significantly mediated through agency attributions (see Table 1). That is, as the harmfulness of the agent increases the target is attributed less agency, which in turn predicts the target's lower moral standing.

Table 1

Results of mediation analyses from Studies 1-4.



Indirect Effect, ab (CI)	SE	Direct Effect, f (*p*-value)
Study 1: Benevolent (= 1) vs. Neutral (= 2) vs. Harmful (= 3)		
-.48 (-.85, -.02)	.21	-.63 (.03)
Study 2a: Benevolent (= 1) vs. Neutral (= 2) vs. Harmful (= 3)		
-.91 (-1.33, -.49)	.21	-.22 (.42)
Study 2b: Benevolent (= 1) vs. Neutral (= 2) vs. Harmful (= 3)		
-.19 (-.44, -.06)	.09	-.64 (.001)
Study 3: Less Harmful (= 1) vs. Moderately Harmful (= 2) vs. Highly Harmful (= 3)		
-.08 (-.25, -.01)	.06	-.31 (.09)
Study 4: Neutral (= 1) vs. White-Collar (= 2) vs. Violent (= 3)		
-.82 (-1.23, -.50)	.18	-.36 (.18)

Note. All analyses used a bootstrapping procedure (5,000 resamples) to construct 95% bias-corrected confidence intervals; bold indirect paths have CIs with values non-overlapping with zero.

2.3 Discussion

The results of Study 1 were consistent with the perspective coming from dehumanization theory, but contradicted the predictions made by moral typecasting theory. Harmful agents were

attributed the lowest levels of agency, followed by the neutral, non-offending agent, while the highest levels of agency were attributed to the benevolent agent. Moreover, agent condition had a significant indirect effect on the perceived moral standing of the agent via agency ratings. In Studies 2a-2b we sought to rule out an alternative explanation (see Footnote 4) that might possibly account for the denial of agency to harmful agents, and assessed agency via an alternate operationalization (Study 2a).

3. Study 2: Ruling Out an Alternative Explanation and Testing an Alternate Operationalization of Agency

Study 2 had three objectives. First, we sought to rule out likeability as a possible explanation for why harmful agents are denied agency. Harmful agents are generally disliked (see Kozak, Marsh, & Wegner, 2006; Waytz & Epley, 2012) and this might account for why they are attributed lower agency, rather than specifically because of their being harmful. In Study 2a we sought to show that likeability cannot completely account for the effects of harmfulness on agency attribution by statistically controlling for perceptions of the target's likeability. In Study 2b, we addressed the issue of likeability experimentally. Second, we sought to show, in Study 2a, that the denial of agency to harmful agents is not limited to the operationalization of agency in terms of mind attribution (Gray et al., 2007), but extends also to other definitions of agency related to activity level and tenacity. Finally, Study 1 was methodologically limited insofar as the agents in the benevolent and harmful vignettes were described as performing a *specific* action while the activity of the agent in the neutral vignette was more abstract. To more tightly standardize the activity level of the agents in Study 2 all three targets were described as performing a specific action.

3.1 Study 2a

3.1.1 Participants

We recruited a new sample of 96 adults (62 male; 34 female; M_{age} = 33.39 years, SD = 10.37) from the same web service as before, excluding individuals who participated in Study 1 and restricting the sample to U.S. residents. Participants were compensated $.30 and the average completion time was three minutes and twenty-six seconds.

3.1.2 Materials and procedures

Participants were randomly assigned to one of the three agent conditions (harmful; neutral; benevolent). In all three conditions, participants first read some filler information about the target: "John is a 29 year old man with brown hair and brown eyes. He works as a taxi driver in Miami." Then they read additional information, which varied depending on the agent condition they were assigned to. In the harmful agent condition, participants read that at night, when John is off work, he likes to spend his time collecting stray cats and dogs to torture and experiment on in his basement. In the benevolent agent condition, John was presented as liking to spend his time helping out with Meals on Wheels to deliver food to the needy. Finally, participants in the neutral agent condition were told that John likes to spend his time practicing guitar and writing music (see Appendix B for full vignettes).

Following the agent manipulation, participants completed the same measures of harmfulness (α = .97), agency (α = .89), and moral standing (α = .96) as in Study 1. Agreement with these items was rated using the same scale sizes and anchors as in Study 1. Perceptions of the target's likeability was assessed in terms of "how much do you like John" (on a 1-7 scale, *Not at all* to *Very much*). In addition to administering the same measure of agency as in Study 1, perceptions of the target's agency was assessed using six "activity" traits: *tenacious, willful, potent, vigorous, active, energetic* (α = .76; all on a 1-7 scale, *Strongly disagree* to *Strongly*

agree). No other measures were collected aside from basic demographics. All participants were debriefed and paid.

3.2 Results

3.2.1 Manipulation check

The harmfulness manipulation produced a significant effect in the expected direction, $F(2,93) = 48.56$, $p < .001$, $\eta^2_p = .511$. The target was rated as more harmful in the harmful agent condition ($M = 5.85$, SD = 1.80) than either in the neutral agent condition ($M = 2.60$, SD = 1.20), $p < .001$, or in the benevolent agent condition ($M = 2.37$, SD = 1.71), $p < .001$, whereas the harmfulness of the target in the benevolent and neutral agent conditions was rated equally low, $p = .56$.

3.2.2 Likeability

There was a significant main effect of agent condition on likeability, $F(2,93) = 77.13$, $p < .001$, $\eta^2_p = .624$. John was perceived as less likeable in the harmful agent condition ($M = 1.46$, SD = 1.25) than in the neutral ($M = 5.00$, SD = 1.28), $p < .001$, and benevolent agent conditions ($M = 5.52$, SD = 1.75), $p < .001$, whereas the target was rated equally likeable in the benevolent and neutral agent conditions, $p = .16$. Collapsing across agent condition, ratings of harmfulness and likeability were strongly correlated, $r(94) = -.90$, $p < .001$.

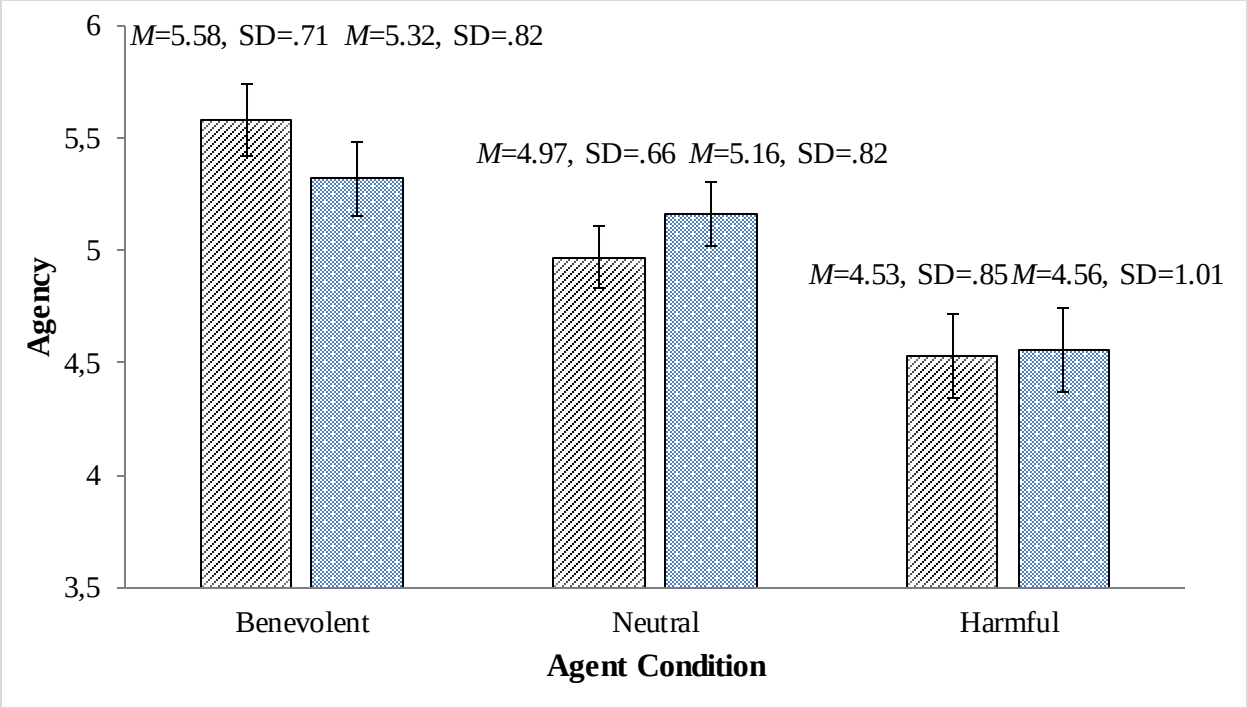
3.2.3 Main analysis and mediation

Collapsing across agent condition, agency and likeability were strongly correlated when operationalized in terms of cognitive agency traits, $r(94) = .81$, $p < .001$, and moderately correlated when operationalized in terms of activity agency traits, $r(94) = .29$, $p = .004$. Three one-way ANCOVAs were conducted on each of the agency indices (cognitive ability and activity traits) and moral standing index, with perceptions of the target's likeability as a covariate

(also see Footnote 5). There was a significant main effect of agent condition on the mind perception agency index used in Study 1, $F(2,92) = 3.41$, $p = .037$, $\eta^2_p = .069$ (see Figure 2 for means and SDs). John was attributed lower levels of agency in the harmful agent condition than in the neutral, $p = .027$, and benevolent agent conditions, $p = .012$, whereas John was attributed similar levels of agency in the neutral and benevolent agent conditions, $p = .43$. Paralleling the results with the cognitive agency traits there was a significant main effect of agent condition on the activity agency index, $F(2,92) = 8.06$, $p < .001$, $\eta^2_p = .149$. Consistent with a dehumanization effect, John was attributed lower levels of agency in the harmful agent condition than in the benevolent agent condition, $p < .001$, and marginally lower levels of agency than in the neutral agent condition, $p = .097$, with the highest levels of agency in the benevolent agent condition (see Figure 2 for means and SDs). Finally, no main effect emerged of agent condition on moral standing ratings, $F(2,92) = 1.30$, $p = .28$ (also see Footnote 6).

Figure 2

Attributions of agency to the human target by agent condition (Study 2a). Stripe bars = cognitive agency traits; dotted bars = activity agency traits. Error bars ± 1 SE.



Similar to Study 1, we used the PROCESS macro to conduct a mediation analysis. We used a bootstrapping procedure (5,000 resamples) to construct bias-corrected confidence intervals. The bootstrapping procedure showed that agent condition had a significant indirect effect on judgments of the agent's moral standing via attributions of agency (see Table 1) operationalized in terms of cognitive ability (also see footnote 7). That is, as the harmfulness of the agent increased the target was attributed less agency, which in turn predicted the target's lower moral standing.

3.3 Discussion

The results of Study 2a were again consistent with a dehumanization perspective, but contradicted the predictions made by moral typecasting theory. Harmful agents were attributed the lowest levels of agency, while the highest levels of agency were largely attributed to the benevolent agent, even when statistically controlling for the agent's likeability. These findings

suggest that likeability cannot completely account for the effect of harmfulness on agency attribution. Moreover, denial of agency to harmful agents occurred for both operationalizations of agency, whether agency was defined using mind perception items or activity/tenacity items. Additionally, agent condition again had a significant indirect effect on the perceived moral standing of the agent via agency ratings. In Study 2b we sought to experimentally disentangle harmfulness and likeability, as much as possible, to test whether the effects of harmfulness on agency are exclusively mediated through likeability.

3.4 Study 2b

3.4.1 Participants

We recruited a new sample of 90 adults (61 male; 29 female; M_{age} = 31.90 years, SD = 10.45) from the same web service as before, excluding individuals who participated in the previous studies or took part in the same study twice, as well as restricting the sample to U.S. residents. Participants were compensated $.50 and the average completion time was eight minutes and 47 seconds.

3.4.2 Materials and procedures

Three vignettes were devised to tease apart harmfulness and likeability: harmful/likeable; neutral/likeable; benevolent/unlikeable. Participants were randomly assigned to one of the three vignettes. All participants first read about a man named David, who is a vegetarian. Then they read additional information about the activity of the target involving their harming or helping others, and, independently, their likeability, which varied by condition (see Appendix B for full vignettes). Following the agent manipulation, participants completed a measure of harmfulness (six traits: *harmful, mean, hostile, peaceful, gentle, caring* [positive traits reverse coded]; α = .96), and the same agency (α = .88) and moral standing (α = .91) measures as in the previous

studies, with agreement with these items rated using the same scale sizes and anchors. The likeability of the target was assessed with the item "David is likeable" rated on a 1-7 scale (*Not at all* to *Very much so*). No other measures were collected aside from basic demographics. All participants were debriefed and paid.

3.5 Results

3.5.1 Manipulation check

The harmfulness manipulation produced a significant effect in the expected direction, $F(2,87) = 93.28$, $p < .001$, $\eta^2_p = .682$. The target was rated as more harmful in the harmful/likeable agent condition ($M = 5.65$, SD = 1.10) than either in the neutral/likeable agent condition ($M = 2.39$, SD = .79), $p < .001$, or in the benevolent/unlikeable agent condition ($M = 2.79$, SD = 1.20), $p < .001$, whereas the harmfulness of the target in the benevolent/unlikeable and neutral/likeable agent conditions was rated equally low, $p = .14$.

3.5.2. Likeability

There was a significant main effect of agent condition on likeability, $F(2,86) = 29.84$, $p < .001$, $\eta^2_p = .407$. David was seen as less likeable in the harmful/likeable agent condition ($M = 2.58$, SD = 1.43), compared to the neutral/likeable ($M = 5.32$, SD = 1.15), $p < .001$, and benevolent/unlikeable agent conditions ($M = 3.76$, SD = 1.76), $p = .003$. That the harmful/likeable target was rated less likeable than the benevolent/unlikeable target likely reflects the contribution of harmfulness to likeability (collapsing across condition, harmfulness and likeability were again negatively correlated, $r(88) = -.76$, $p < .001$). Critically, however, David was seen as more likeable in the neutral/likeable agent condition than in the benevolent/unlikeable agent condition, $p < .001$. This shows that harmfulness and likeability are, to an extent, separable constructs.

3.5.3 Main analysis and mediation

Two one-way ANOVAs were conducted on the agency index and moral standing index. There was a significant main effect of agent condition on the agency index, $F(2,86) = 10.75$, $p < .001$, $\eta^2_p = .200$. David was attributed lower levels of agency in the harmful/likeable agent condition ($M = 4.44$, SD = 1.45) than in the neutral/likeable ($M = 5.65$, SD = .75), $p < .001$, and benevolent/unlikeable agent conditions ($M = 5.29$, SD = .83), $p = .004$, whereas David was attributed similar levels of agency in the neutral/likeable and benevolent/unlikeable agent conditions, $p = .20$.

Paralleling the results with the agency index, there was a significant main effect of agent condition on the moral standing index, $F(2,86) = 17.64$, $p < .001$, $\eta^2_p = .291$. David was attributed lower levels of moral standing in the harmful/likeable agent condition ($M = 4.46$, SD = 1.54) than in the neutral/likeable ($M = 6.03$, SD = .82), $p < .001$, and benevolent/unlikeable agent conditions ($M = 6.06$, SD = 1.15), $p < .001$, whereas David was attributed similar levels of moral standing in the neutral/likeable and benevolent/unlikeable agent conditions, $p = .92$.

To test the individual contributions of harmfulness and likeability on attributions of agency, we conducted a linear regression with the two factors entered as simultaneous predictors of agency. Replicating Study 2a results, harmfulness emerged as a significant predictor of agency attributions, $\beta = -.51$, $t(86) = -3.70$, $p < .001$, likeability did not, $\beta = .07$, $t(86) = .48$, $p = .63$. Finally, we used the PROCESS macro to conduct mediation analyses. We used a bootstrapping procedure (5,000 resamples) to construct bias-corrected confidence intervals. The bootstrapping procedure showed that agent condition had a significant indirect effect on judgments of the agent's moral standing via attributions of agency (see Table 1). That is, as the harmfulness of the agent increased the target was attributed less agency, which in turn predicted

the target's lower moral standing. Importantly, when we conducted a multiple-mediation analysis with harmfulness and likeability ratings entered simultaneously as mediators of the effect of agent condition on attributions of agency, harmfulness emerged as a significant mediator of agency attributions, indirect effect = -.59 (95% bias corrected CIs [-1.06, -.23]), while likeability did not, indirect effect = -.01 (95% bias corrected CIs [-.18, .11]).

3.6 Discussion

The results of Study 2b were consistent with the results from Study 2a, but this time likeability was experimentally manipulated independent from harmfulness. In this study harmfulness again significantly reduced attributions of agency, independent from likeability, and perceptions of harmfulness mediated the effect independent of likeability as well. Finally, harmfulness again had a significant indirect effect on the perceived moral standing of the agent via agency ratings. Together the findings of Study 2 suggest that the effect of harmfulness on perceptions of agency is not attributable exclusively to the likeability of the agent. In Study 3 we sought to replicate our findings using a different agent type – a cigarette corporation – and even greater experimental control over the activity level of the agent.

4. Study 3: The Agency of Harmful Companies

Past research has shown that corporations are often anthropomorphized and can be perceived as intentional agents (Cohen, 2014; Kervyn, Fiske, & Malone, 2012) with distinct personalities (Aaker, 1997); furthermore, consumers often exhibit levels of emotional attachment towards specific companies and brands (Thomson, MacInnis, & Park, 2005). In Study 3 we capitalized on this tendency to treat corporations as intentional agents and sought to replicate the findings of Studies 1, 2a, and 2b using an agentive, nonhuman entity. We provided participants information about the same agent, a manufacturer of cigarettes, while modifying only the nature

of the manufactured product: the harmfulness of the cigarette. This procedure allowed us to tightly standardize the activity of the three corporations, while varying only the level of harm caused by the agent. To prevent the possibility of familiarity and loyalty effects to an existing company, we used a fictional corporate brand.

4.1 Method

4.1.1 Participants

We recruited a new sample of 98 adults (60 male; 37 female; 1 undisclosed; M_{age} = 33.18 years, SD = 10.29) from the same web service as before, excluding individuals who participated in the previous studies and restricting the sample to U.S. residents. Participants were compensated $.25 and the average completion time was two minutes and thirty-three seconds.

4.1.2 Materials and procedures

Participants were randomly assigned to one of the three agent conditions (highly harmful; moderately harmful; less harmful). In the *highly harmful agent* condition, participants read about Initrode, a cigarette company that makes harsh cigarettes that lead to a higher incidence of lung cancer. In the *less harmful agent* condition, Initrode was presented as a cigarette company that makes light cigarettes that lead to a lower incidence of lung cancer. Finally, participants in the *moderately harmful agent* condition were provided the same information about Initrode, the cigarette company, without mentioning the degree of harmfulness caused by the cigarette or the resulting incidence of lung cancer (see Appendix B for full vignettes).

One thing to note about the moderately harmful and less harmful agent conditions is that the agents in these conditions are arguably still causing some harm—though *relatively less harm* than the highly harmful agent—since all three companies are producing cigarettes. This is

different from Studies 1, 2a, and 2b where the neutral agent was a non-offender and the benevolent agent produced exclusively beneficial outcomes.

Following the agent manipulation, participants completed the same measures of harmfulness ($\alpha = .90$), agency ($\alpha = .86$), and moral standing ($\alpha = .90$) as in Study 1, with "Initrode" as the target instead of "John." Agreement with these items was rated using the same scale sizes and anchors as in the previous studies. No other measures were collected aside from basic demographics. All participants were debriefed and paid.

4.2 Results

4.2.1 Manipulation check

The harmfulness manipulation was successful. There was a significant difference in the amount of harmfulness perceived across the three companies, $F(1,95) = 15.20$, $p < .001$, $\eta^2_p = .242$. The highly harmful company was rated as more harmful ($M = 5.26$, SD = 1.41) than either the moderately harmful company ($M = 4.07$, SD = 1.07), $p = .001$, or the less harmful company ($M = 3.59$, SD = 1.32), $p = .001$, whereas the harmfulness of the less and moderately harmful companies was rated equally low, $p = .15$, though there was a non-significant trend to rate the less harmful company less harmful than the moderately harmful company.

4.2.2 Main analysis and mediation

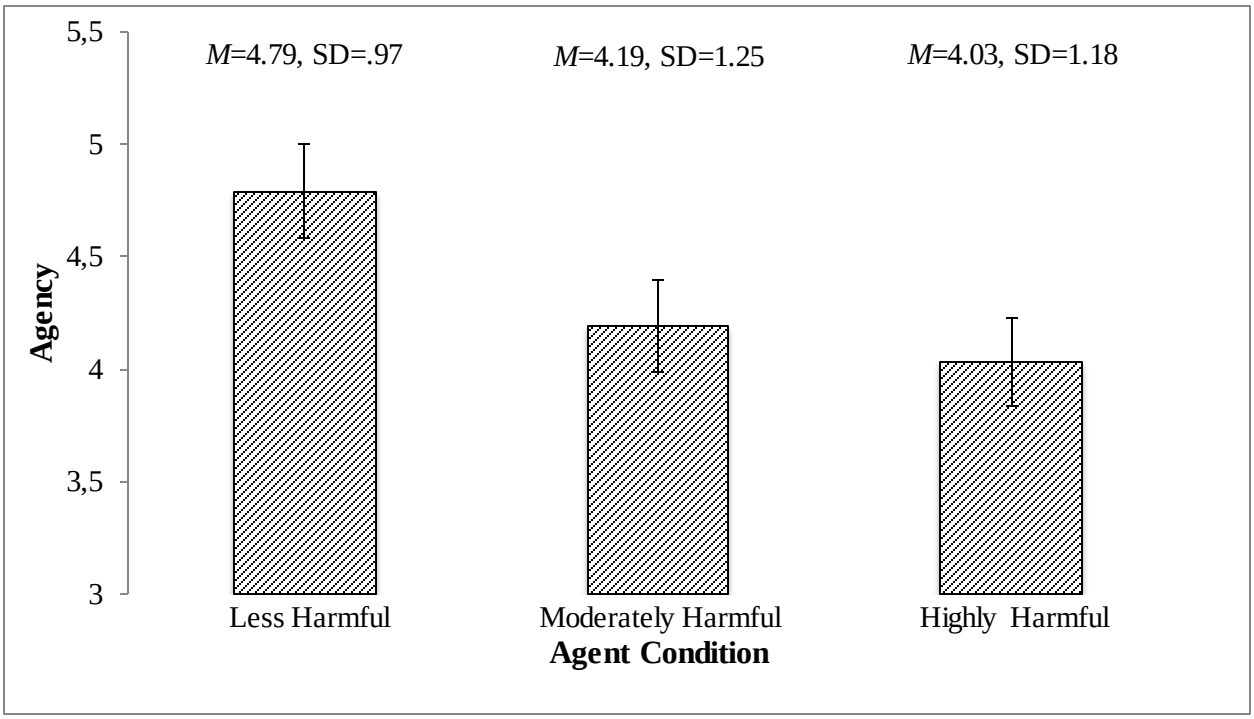
Two one-way ANOVAs were conducted on the agency index and moral standing index. There was a significant main effect of condition on the agency index, $F(1,95) = 3.81$, $p = .026$, $\eta^2_p = .076$. Pairwise comparisons showed that the less harmful company was attributed higher levels of agency than the moderately harmful company, $p = .04$, and the highly harmful company, $p < .003$ (see Figure 3). The highly harmful company and moderately harmful company did not significantly differ in their perceived agency, $p = .56$. There was a significant

main effect of agent condition on moral standing ratings, $F(1,95) = 3.84$, $p = .025$, $\eta^2_p = .076$. Pairwise comparisons showed that the highly harmful company was attributed lower moral standing ($M = 4.07$, SD = 1.79) than the less harmful company ($M = 4.81$, SD = 1.13), $p = .04$, and the moderately harmful company ($M = 4.96$, SD = 1.11), $p = .01$. The less harmful company and moderately harmful company did not significantly differ in their attributed moral standing, $p = .66$.

Similar to Studies 1-2b, we used the PROCESS macro to conduct a mediation analysis. We used a bootstrapping procedure (5,000 resamples) to construct bias-corrected confidence intervals. The bootstrapping procedure showed that agent condition had a significant indirect effect on judgments of the agent's moral standing via attributions of agency (see Table 1). The highly harmful and moderately harmful corporations were each attributed less agency than the less harmful corporation, and this decrease in agency predicted levels of moral standing.

Figure 3

Attributions of agency to the cigarette company by agent condition (Study 3). The less harmful cigarette company produced light cigarettes. Error bars ± 1 SE.



4.3 Discussion

Study 3 replicated the results of Studies 1, 2a, and 2b using a nonhuman entity as the harmful agent. The findings were again consistent with a dehumanization perspective and inconsistent with predictions made by MTT. The cigarette corporation that produced highly harmful cigarettes was denied agency to a greater extent than the company that produced less harmful, light cigarettes (with the moderately harmful cigarette company denied agency to a similar extent since it arguably still causes considerable harm by virtue of being a cigarette manufacturer). As a consequence, the highly harmful corporation was attributed the lowest level of moral standing.

5. Study 4: The Agency and Humanity of Harmful Agents

In Study 4, we sought to replicate the results of Studies 1-3 using a new set of vignettes taken from prior research, while also expanding our assessment of agency beyond the cognitive traits used in the mind perception literature. According to several groups of researchers (e.g.,

Bastian et al., 2011; Haslam et al., 2008; Haslam & Loughnan, 2014; Waytz et al., 2010), there is significant overlap in the agency traits utilized by Gray et al. (2007) and the Human Uniqueness traits postulated by Haslam (2006) in his two dimensions of humanness (human nature vs. human uniqueness) model of dehumanization. Human uniqueness traits are thought to distinguish humans from animals, and involve attributes related to rationality, civility, and cultural refinement, whereas human nature traits are traits that are thought to typify human beings and distinguish us from inanimate objects—traits such as emotionality, responsiveness, and warmth (see Haslam, 2006; Haslam et al., 2008; Loughnan & Haslam, 2007). Theorists from both the mind perception literature (e.g., Waytz et al., 2010) and the dehumanization literature (e.g., Haslam & Loughnan, 2014) have speculated that human uniqueness traits have much in common with agency traits, though arguably the construct of human uniqueness may be somewhat broader. To our knowledge the only study to date that has attempted to assess the mind dimension of agency and the humanness dimensions of human uniqueness within the same study was conducted by Bastian et al. (2011). However, in their study blame and praise (potential consequences of perceiving agency) were used as proxies for agency, rather than assessing aspects of agency directly, much like in Gray and Wegner's (2009) studies. The authors found that judgments of blame were influenced more by a target's level of human uniqueness traits (e.g., rationality) than a target's level of human nature traits (e.g., emotional warmth). Agency and human uniqueness may be related constructs, but no study to date, as far as we are aware, has included assessments of agency traits and humanness traits within the same study. Study 4 helps fill this gap.

5.1 Method

5.1.1 Participants

Eighty-four adult participants (38 male; 46 female; M_{age} = 35.89 years, SD = 11.58) located in the US participated in the study via the same web service as before; individuals who participated in the previous studies were excluded. Participants were compensated $.40 and the average completion time was four minutes and twenty five seconds.

5.1.2 Materials and procedures

Participants were randomly assigned to one of the three agent conditions (violent offender; white collar criminal; neutral agent). The harmful agent vignettes were adapted from Bastian et al. (2013). We did not include a benevolent agent in this study, but instead included two different levels of harmfulness. In the violent offender and white-collar criminal conditions, the same person was described as having committed a violent or a white-collar crime. In the neutral agent condition, we simply described an average person who was a teacher (see Appendix B for full vignettes).

The dependent measures were identical to those used in Studies 1-3, with the addition of the denial of Human Nature and Human Uniqueness measure taken from Bastian and Haslam (2010). The measure is comprised of eight items, four items for each subscale (see Appendix A for full list of items). The denial of Human Nature subscale includes items such as: "I felt like the person in the story was emotional, like he was responsive and warm" [*reverse scored*] (α = .87). The second 4-item denial of Human Uniqueness subscale included items such as: "I felt like the person in the story was rational and logical, like he was intelligent" [*reverse scored*] (α = .90). Agreement with these items was rated on a 1 (*Not at all*) to 7 (*Extremely so*) scale. Although the two subscales have been treated separately in past research on intergroup dehumanization (Bastian & Haslam, 2010), the scales formed a single dimension in studies by Bastian et al. (2013), which focused on harmful offenders. Consistent with this prior research, a

principle components analysis with Varimax rotation produced a one-factor solution, explaining 69.52% of the total variance; the eigenvalue (5.56) for the first factor was the only value exceeding the conventional cut-off of 1. For this reason, and because the two subscales when combined exhibited a strong internal reliability ($\alpha = .93$), they were aggregated into a single index of *denial of humanness*. As argued by Bastian et al. (2013), the single-factor solution most likely reflects the broad manner in which harmful agents are denied humanness.

5.2 Results

5.2.1 Manipulation check

There was a large main effect of agent condition on harmfulness ratings, $F(1,84) = 150.43$, $p < .001$, $\eta^2_p = .788$. The violent offender was rated as more harmful ($M = 6.41$, SD = .64) than the white-collar criminal ($M = 5.26$, SD = 1.03), who was rated more harmful than the neutral agent ($M = 2.27$; SD = 1.09); all pairwise comparisons, ps $< .001$.

5.2.2 Main analysis and mediation

First, a one-way ANOVA was conducted on the agency index. There was a significant main effect of agent condition on agency ratings, $F(1,77) = 44.04$, $p < .001$, $\eta^2_p = .534$, that was again consistent with a dehumanization effect. The most harmful agent (violent offender) was attributed the lowest levels of agency ($M = 3.60$, SD = 1.13), lower than both the white-collar criminal ($M = 4.14$, SD = .92), $p = .03$, and the neutral agent ($M = 5.86$, SD = .60), $p < .001$. Additionally, the white-collar criminal was attributed less agency than the neutral agent, $p < .001$. Thus, the results regarding agency attribution were largely consistent with Studies 1-3.

We also conducted two separate one-way ANOVAs on the denial of humanness and moral standing indices. There was a significant main effect of agent condition on the denial of humanness index, $F(1,80) = 110.24$, $p < .001$, $\eta^2_p = .734$, consistent with the findings of Bastian

et al. (2013). The violent criminal was dehumanized (M = 5.65, SD = .83) more so than the white-collar criminal (M = 5.00; SD = 1.08), p = .009, and the neutral agent (M = 2.26, SD = .72), p < .001. Additionally, the white-collar criminal was dehumanized more than neutral agent, p < .001. We also found a significant main effect of agent condition on moral standing ratings, $F(1,81) = 21.51$, $p < .001$, $\eta^2_p = .353$. The neutral agent was attributed higher moral standing (M = 6.33, SD = .73) than the white-collar criminal (M = 4.13, SD = 1.57), p < .001, and the violent criminal (M = 3.94, SD = 1.92), p < .001. The violent criminal and the white-collar criminal did not significantly differ in their attributed moral standing, p = .65.

Similar to Studies 1-3, we used the PROCESS macro to conduct a series of mediation analyses, first using agency and denial of humanness as separate mediators and secondly treating agency and denial of humanness as simultaneous mediators (ratings of agency and denial of humanness were significantly correlated, $r(80) = -.83$, $p < .001$, tolerance = .303, VIF = 3.30; the factor analysis results including agency and denial of humanness items may be found in Supplementary Materials). Each analysis used a bootstrapping procedure (5,000 resamples) to construct bias-corrected confidence intervals. The bootstrapping procedure provided support for the expected mediation model: there was a significant indirect effect of agent condition (1 = neutral; 2 = white collar; 3 = violent) on moral standing mediated through agency (see Table 1). We also found support for the expected mediation model, indirect effect = -.99 (95% bias corrected CIs [-1.51, -.44]), when denial of humanness was entered as the mediator. When agency and denial of humanness were treated as simultaneous mediators within a multiple-mediation analysis, the indirect effect of agent condition on judgments of the agent's moral standing operated via agency attributions, ab_{agency} = -.67 (95% bias corrected CIs [-1.22, -.29]),

but not through the denial of humanness variable, $ab_{dehumanization}$ = -.33 (95% bias corrected CIs [-.98, .37]).

5.3 Discussion

We yet again replicated the finding that agents are denied agency, not attributed agency, as a consequence of their harmful actions. Study 4 replicated this dehumanization pattern with yet another set of individuals, this time with two gradations of harmful agents: violent offender and white-collar criminal. Study 4 also showed that it is the denial of agency in particular (as opposed to the denial of humanity broadly construed) that mediates the effect that perceived harmfulness has on judgments of the agent's moral standing.

6. Study 5: Inferential and Motivational Causes of Agency Denial

Thus far we have shown that the inferences perceivers form about harmful agents are consistent with a dehumanization process and inconsistent with the moral typecasting perspective. However we have yet to specify the reasons why perceivers deny harmful agents agency in the first place. In Study 5 we sought to investigate two potential mechanisms of agency denial: (1) that perceivers *infer* low levels of agency (rationality, self-control, etc.) from an agent's harmful behavior; and (2) that perceivers are *motivated* to view the harmful agent as lacking agency in order to help legitimate or justify punishing the agent. To test for these two potential mechanisms, in Study 5 we independently manipulated in a 2 x 2 between-subjects design: (1) the perceived rationality of the offense (rational vs. irrational harm), to test the inferential hypothesis; and (2) motivations to see the agent punished (agent harms a loved one vs. harms a criminal), to test the motivational hypothesis. We predicted that a rational harmful agent would be attributed more agency than an irrational harmful agent, and that an agent that causes

harm to a loved one (and thus generates greater retributive justice motivations) would be attributed less agency than an agent that causes harm to a criminal.

6.1 Method

6.1.1 Participants

A new sample of 114 adult participants (64 male; 49 female; 1 undisclosed; M_{age} = 37.71 years, SD = 13.48) located in the US participated in the study via the same web service as before, excluding individuals who participated in the previous studies. Participants were compensated $.50 and the average completion time was five minutes and four seconds.

6.1.2 Materials and procedures

Design. We used a 2 (rationality of offense: high vs. low) x 2 (motivation to punish offender: high vs. low) between-subjects design. All participants read the same initial information about the harmful agent: "John is a 23 year old male. He is about six feet tall, with brown hair and blue eyes. John has taken to crime."

Manipulations. Depending on which motivation condition participants were randomly assigned to, they were instructed to further imagine that one day John mugs the participants' mother (high motivation) or a gang member (low motivation). Depending on which rationality condition participants were randomly assigned to, John was described as performing the harmful act because he needs the money to feed himself (high rationality) or for the fun of it (low rationality). See Appendix B for full vignettes by condition.

Manipulation checks. Participants completed two manipulation checks. First, as a check on the perceived rationality manipulation, participants rated how *reasonable, rational, sensible,* and *understandable* (1 = *Not at all,* 7 = *Extremely,* α = .92) were John's actions. Second, as a check on the motivation to punish manipulation, they rated how strongly they wanted *to retaliate*

against John for the incident, *to take revenge on John*, and *to get even with John* ($\alpha = .98$), all on a 1-7 scale (*Strongly disagree* to *Strongly agree*).

Main dependent measures. The measures of agency, dehumanization, and moral standing were identical to those used in Study 4. We again submitted the denial of Human Nature and the denial of Human Uniqueness subscales to a principle components analysis with Varimax rotation. The analysis produced a two-factor solution, explaining 66.75% of the total variance, with the eigenvalues for the two factors equal to 4.12 and 1.22. However, these two factors largely represented a difference between reverse-coded and non-reverse-coded items and not differences in denial of human uniqueness and denial of human nature. Furthermore, a parallel analysis suggested that the second factor should be disregarded as a separate factor. For this reason, and because the two subscales exhibited strong internal reliability ($\alpha = .85$), the eight items were aggregated into a single index of *denial of humanness* similar to Study 4.

We also included in Study 5 for exploratory purposes the patiency or "experience" items taken from Gray et al. (2007), with experience/patiency defined broadly in terms of experiential cognitive capacities, such as the capacity for consciousness and to experience various emotions (e.g., pleasure, pain) and drive states (e.g., hunger). Eleven patiency traits were included in total ($\alpha = .91$; see Appendix A for a full list of items). Agreement with these items was rated on a 1 (*Strongly disagree*) to 7 (*Strongly agree*) scale. In Gray et al.'s (2007) initial study a factor analytic approach was adopted using Varimax rotation that forced orthogonality between measures of agency and patiency; however, the raw correlations between the factors were not reported. Reanalysis of their data set, as reported by Piazza et al. (2014), revealed quite high positive correlations between these dimensions ($r = .90$). Other research using wider sets of non-human animal targets have revealed similarly high levels of overlap between Gray et al.'s agency

and experience items (see Bastian, Loughnan, Haslam, & Radke, 2012, Study 1). In the present study, the agency and patiency indices were highly correlated, $r(108) = .62$, $p = .001$. Since our focus in the present study was on agency in particular, we created separate agency and patiency indices and assessed these measures separately in the analyses below. The results for patiency are reported in Supplementary Materials. Analyses with the patiency items and the agency and patiency items combined into a single index may be found in Supplementary Materials as well. No other measures were collected aside from basic demographics.

6.2 Results

6.2.1 Manipulation checks

There was a main effect of the rationality manipulation on rationality ratings, $F(1,112) = 8.03$, $p = .005$, $\eta^2_p = .067$. The offense was rated more rational in the high rationality condition ($M = 2.07$, SD = 1.35) than in the low rationality condition ($M = 1.45$, SD = .96). The rationality manipulation did not affect motivations to punish the offender, $F(1,111) = .53$, $p = .47$, $\eta^2_p = .005$, nor did it interact with the motivation manipulation to affect rationality ratings, $F(1,110) = 1.45$, $p = .23$, $\eta^2_p = .013$, or ratings to see the agent punished, $F(1,109) = .01$, $p = .91$, $\eta^2_p = .000$. Thus, the rationality manipulation had its intended effect, and this effect was unique to rationality ratings.

There was a main effect of retribution motivation on ratings to see the agent punished, $F(1,111) = 35.75$, $p < .001$, $\eta^2_p = .244$. Participants' motivations to see the agent punished were higher in the high motivation condition ($M = 5.33$, SD = 1.84) than in the low motivation condition ($M = 3.34$, SD = 1.70). However, the motivation to punish manipulation also incidentally affected perceptions of rationality of the offense, $F(1,112) = 4.85$, $p = .030$, $\eta^2_p = .042$. The agent whom participants were more motivated to see punished was also perceived to

be somewhat less rational (M = 1.51, SD = 1.07) than the agent whom participants were less motivated to punish (M = 2.00, SD = 1.29). This incidental effect may reflect a motivated attribution process (e.g., see Alicke, 2000), whereby motivations to punish an agent causes perceivers to derogate the target in other respects. Thus, while our motivation manipulation was successful, and its intended effect was quite large, it also had an incidental effect on rationality ratings.

6.2.2. Preliminary correlations of agency and denial of humanness

As in Study 4, ratings of agency and denial of humanness were significantly correlated, $r(109) = -.44$, $p < .001$ (the factor analysis results including agency and denial of humanness items may be found in Supplementary Materials). Multicollinearity between these variables was not an issue (tolerance = .795, VIF = 1.26); thus, we treated agency and denial of humanness as separate dependent measures.

6.2.3 Main analysis and mediation

We conducted four separate 2 x 2 ANOVAs on the agency, patiency, denial of humanness and moral standing indices.

Agency. There was a main effect of rationality condition on agency ratings, $F(1,108) = 4.40$, $p = .038$, $\eta^2_p = .039$. As predicted by the inferential hypothesis, the less rational agent was attributed lower levels of agency (M = 3.26, SD = 1.41) than the more rational agent (M = 3.79, SD = 1.21). However, there was no significant main effect of motivation to punish condition on agency ratings, $F(1,108) = .14$, $p = .71$, $\eta^2_p = .001$. Additionally, there was no interaction effect of rationality and motivation to punish on agency ratings, $F(1,106) = 1.99$, $p = .17$, $\eta^2_p = .018$. Ratings of the perceived rationality of the offense were also significantly correlated with agency

ratings, $r(110) = .24$, $p = .013$. Ratings of agency and participants' motivations to see the agent did not correlate, $r(109) = -.049$, $p = .61$.

Denial of humanness. There was a significant main effect of rationality condition on denial of humanness ratings, $F(1,109) = 6.04$, $p = .016$, $\eta^2_p = .053$, that was consistent with the inferential hypothesis. The less rational agent was dehumanized more ($M = 6.14$, SD = .87) than the more rational agent ($M = 5.71$, SD = .95). Additionally, there was a significant main effect of motivation to punish on denial of humanness ratings, $F(1,109) = 7.18$, $p = .008$, $\eta^2_p = .062$, that was consistent with the motivational hypothesis. The agent whom participants were more motivated to see punished was dehumanized more ($M = 6.15$, SD = .87) than the agent whom participants were less motivated to see punished ($M = 5.69$, SD = .94). There was no interaction effect of rationality and motivation to punish on denial of humanness ratings, $F(1,107) = .23$, $p = .633$, $\eta^2_p = .002$.

Moral standing. There was a significant main effect of rationality condition on moral standing ratings, $F(1,108) = 7.72$, $p = .006$, $\eta^2_p = .067$. The less rational agent was attributed lower moral standing ($M = 3.80$, SD = 1.83) than the more rational agent ($M = 4.63$, SD = 1.54). There was also a significant main effect of motivation to punish condition on moral standing ratings, $F(1,108) = 16.10$, $p < .001$, $\eta^2_p = .130$. The agent whom participants were more motivated to see punished was attributed lower moral standing ($M = 3.57$, SD = 1.90) than the agent whom participants were less motivated to see punished ($M = 4.79$, SD = 1.34). There was no interaction effect of rationality and motivation to punish on moral standing ratings, $F(1,108) = 1.16$, $p = .28$, $\eta^2_p = .011$.

Mediation. We used the PROCESS macro to conduct a series of mediation analyses, first using agency and denial of humanness as separate mediators of the significant effects of

rationality and motivation to punish on moral standing (we did not conduct the motivation to punish → agency → moral standing mediation analysis since motivation to punish did not significantly influence agency ratings), and secondly, treating agency and denial of humanness as simultaneous mediators. Each analysis used a bootstrapping procedure (5,000 resamples) to construct bias-corrected confidence intervals. The bootstrapping procedure provided support for the expected mediation model with regards to rationality → agency → moral standing: there was a significant indirect effect of rationality condition on moral standing mediated through agency, indirect effect = -.17 (95% bias corrected CIs [-.52, -.01]). There was a significant indirect effect of offense rationality on moral standing mediated through denial of humanness, indirect effect = -.22 (95% bias corrected CIs [-.52, .04]). We also found support for the indirect effect of motivation to punish on moral standing mediated through denial of humanness, indirect effect = .22 (95% bias corrected CIs [.06, .50]), but not through agency, indirect effect = .05 (95% bias corrected CIs [.-11, .27]).

When agency and denial of humanness were entered as simultaneous mediators within a multiple-mediation analysis, the effect of offense rationality on judgments of moral standing operated both through agency (to an extent), ab_{agency} = -.13 (95% bias corrected CIs [-.47, .00]), and denial of humanness attributions, $ab_{dehumanization}$ = -.17 (95% bias corrected CIs [-.49, -.02]). By contrast, the motivation to punish influenced judgments of the agent's moral standing via denial of humanness attributions, $ab_{dehumanization}$ = .15 (95% bias corrected CIs [.01, .41]), but not through agency attributions, ab_{agency} = .03 (95% bias corrected CIs [-.12, .25]).

6.3 Discussion

Study 5 investigated two separate psychological processes that may help promote the denial of agency to harmful agents: an inferential process and a motivational one. However,

evidence for the inferential process was relatively clearer, and emerged despite the fact that our manipulation of rationality was less potent than our manipulation of retribution motives. Rational offenders were ascribed more agency (and were dehumanized less) than irrational offenders, and attributions of agency mediated the influence perceived rationality had on moral standing judgments. Attributions of humanness also served as an independent mediator of rationality and moral standing. By contrast, the motivation to punish manipulation did not affect attributions of agency, though it did affect attributions of humanness: motivations to punish the offender led to increased dehumanization of the offender, and dehumanization levels mediated the effect of punishment motivations on moral standing judgments. Below we discuss possible explanations for why our measure of humanness was more sensitive to the motivational variable than our measure of agency.

7. General Discussion

7.1 Summary of the present studies

Across six studies, we found evidence consistent with a dehumanization perspective on the way lay people conceptualize the agency of harmful agents, evidence that directly contradicts assumptions made by moral typecasting theory. In contrast with benevolent and neutral (non-offending) agents, harmful agents were consistently attributed *less agency* and, as a consequence, were afforded less moral standing (e.g., granted fewer protections and rights). This occurred across different intentional agent categories (humans and companies), when we removed the confounding influence of agents' likeability (Studies 2a and 2b), and when controlling the specific action of the agent (Study 3). This effect was also observed using two different operationalizations of agency (Study 2a), and across different gradations of harmfulness (Studies 3-4). We also demonstrated that harmful agents are denied agency primarily through the

inferences observers draw from the agent's harmful conduct (and the motivations supporting this conduct), and less because of perceivers' motivations to see the agent punished (Study 5).

Taken together, the current set of experiments seems to establish a clear challenge to the moral typecasting account. While moral typecasting theory (Gray & Wegner, 2009) predicts that harmful agents should be attributed levels of agency on par with benevolent agents, by virtue of being typecasted as moral agents, and levels of agency certainly higher than neutral, non-offending agents, the current findings were more consistent with research on dehumanization (e.g., Bastian et al., 2013), which predicts that harmful agents should be attributed less agency (and overall less humanity traits) than both neutral agents and benevolent agents. The fact that harmful agents are denied agency, rather than ascribed agency, directly contrasts with the idea coming from MTT that individuals who harm or help others are typecasted as moral agents, and thus are ascribed agency-relevant traits (see also Gray, 2010; Gray & Schein, 2012; Gray et al., 2012; Gray & Wegner, 2011).

7.2 Theoretical implications, limitations, and future directions

Our findings have clear implication for the future of moral typecasting theory (Gray & Wegner, 2009). At minimum, our results suggest that a reframing of MTT is in order. One possibility is that people do typecast *some* moral agents but not others. For example, we found that benevolent agents (humans and companies who perform good deeds) were attributed greater agency than non-offending and harmful agents. Thus, the moral typecasting hypothesis may still apply, in a more restrictive manner, to perceivers' conceptualization of *good* moral agents. Another possibility is that the moral typecasting hypothesis may apply when we restrict the definition of agency to intentionality and blame. The problem with such a restructuring however is that it relies on a limited assessment of agency in terms of intentionality and the potential

consequences of perceiving agency (e.g., blame), rather than on a rich and comprehensive assessment of the basic manifestations of agency. Arguably, intentionality is a basic feature of agency. However, the attribution of intentionality is not limited to harmful agents, but is a feature easily conferred on all agents regardless of the content of their actions—harmful, benevolent, neutral, or otherwise (see Waytz et al., 2010). Thus, intentionality seems like a poor foundation for establishing agency as it applies to moral agents.

One possibility is that the moral typecasting framework applies to harmful agents only with regards to the denial of patiency. Nonetheless, the current findings appear to call into question whether it is the typecasting of harmful agents *as agentive* that causes the reduction in patiency, as MTT suggests, or whether the denial of patiency to harmful agents is caused by more basic inferences about their emotional callousness (for a similar argument, see Arico, 2012). Another possibility is that the moral typecasting hypothesis may be limited to a very narrow sense of agency that focuses exclusively on power, or the capacity to exert force. Such an argument might be made on the basis of work from Gray (2010), who found that getting people to think and write about a harmful action performed by a fictional agent led people to exert more physical force in a subsequent weight-holding task, compared to a control condition where participants thought and wrote about doing some work. (Getting people to think and write about a helpful action also led somewhat to increased weight-holding times, compared to control.) The results were interpreted as evidence for a moral transformation effect: doing harm or doing good (or at least writing about harming or helping) empowers people. It is possible that the moral typecasting hypothesis applies in a very narrow sense to feelings of power. Our studies did not assess agency narrowly in terms of power, nor did we assess participants' own feelings of power, but had participants make attributions of agency (broadly defined) of other agents. Therefore,

there are a number of methodological differences between our studies and Gray's (2010) study, which make direct comparisons impossible. Nevertheless, it seems unlikely that perceivers would consistently attribute power to harmful agents, vis-à-vis non-offenders, as it seems clear that, at least in some situations, it takes more power (i.e., self-control) to stop oneself from causing harm than from causing it. Furthermore, in Study 2a we operationalized agency in terms of traits very close to the meaning of power (e.g., "potent," "tenacious") and replicated the dehumanization effect. Moreover, Gray and Wegner (2009, Study 4b) themselves found some evidence that harmful and neutral agents are indistinguishable in terms of attributions of power. Thus, any reorganization of MTT around the concept of power would have to contend with these issues.

Our findings also contribute to work on dehumanization. First, to the best of our knowledge Studies 4-5 are the first to assess Gray et al.'s (2007) mind perception attributes and Haslam's (2006) humanness traits within the same study. Past researchers have theorized that there is great overlap in the mind-based agency traits (e.g., planning, emotion recognition, imagination) employed within the mind perception literature and the uniquely human traits (e.g., rationality, refinement, self-restraint) employed in the dehumanization literature (e.g., see Bastian et al., 2011; Haslam et al., 2008; Haslam & Loughnan, 2014; Waytz et al., 2010). Indeed, in the present studies we found moderate to large correlations between agency traits and humanness traits. It is worth noting however that the correlations were not consistently large and the results of the factor analyses do not support the conclusion that the two constructs are identical. Furthermore, the agency traits seemed to correlate not only with uniquely human traits (traits like rationality that set humans apart from animals), but also the human nature traits (traits like emotionality that set humans apart from machines). Indeed, in our studies both sets of

humanness traits tended to form a single factor, which is consistent with past studies by Bastian et al. (2013) which focused on harmful agents as targets. Harmful agents, it would seem, are not only seen as lacking in rationality, self-restraint, and civility, as well as specific higher order cognitive capacities (e.g., planning, imagination), but also are seen as lacking emotionality, warmth, and other experience-based traits. In short, harmful agents are seen both in animalistic terms (lacking civility, reason, self-restraint) as well as possessing the qualities of machines (unfeeling automatons). As a consequence of this, harmful agents are denied the rights and protections owed to non-offending individuals: as we observed across all five studies, the denial of agency (and other aspects of humanness) to harmful agents mediated judgments of whether the agent was owed moral standing.

Second, the present work highlights two independent mechanisms that help clarify the causal process by which harmful agents are denied agency. First, denial of agency and humanness appear to be rooted in an inferential process whereby perceivers infer a lack of agency by the very fact that the agent inflicts harm on others, in effect reasoning that only an irrational, unrestrained, impulsive, uncivilized individual would behave in such a ruthless manner. We reasoned that if participants infer the lack of agency from the agent's harmful conduct, then manipulations of the agent's rationale for causing harm should exaggerate this process, as participants draw inferences about the harmful agent's underlying agency. Indeed, in Study 5, this is exactly what we found.

The second putative mechanism we tested was moral disengagement, a motivational process whereby observers may deny an offender agency or humanness in order to justify punishing or aggressing towards them (see Bandura, 1999; Castano & Giner-Sorolla, 2006; Leidner et al., 2013; Leidner et al., 2010). To test this idea, in Study 5 we manipulated

participants' motivations to punish the offender, to see if participants would deny the agent agentive qualities even more strongly when motivated to punish the offender. The results regarding this moral disengagement process were less clear. Although participants dehumanized the harmful agent more when motivated to see him punished, they were no more likely to deny the agent agentive traits under increased justice motivations.

We might speculate that motivations to punish the offender failed to influence attributions of agency due to competing motivational forces operating in the process of meting out justice. On the one hand, according to moral disengagement theory, when a person is motivated to harm someone, they must engage in counter-measures to lessen the potential offense of doing so. Dehumanizing the target is one such measure insofar as people feel less guilty about harming someone lacking the qualities befitting a human (Bandura, 1999; Haslam & Loughnan, 2014). On the other hand, the law provides exceptions for individuals who lack agency (e.g., children, mentally impaired) when sentencing criminal acts (Christopher & Pinals, 2010; Fitzgerald, 1962; Hart, 1968; Robinson, 1984). Thus, under motivations to see someone punished, it might be in one's interests to play up an agent's agentive capacities (e.g., level of cognizance). It is quite possible then that inducing motivations to punish a harmful agent causes an interaction of opposing strategies whereby the target is at once denied humanness but ascribed some level of agency to ensure they are properly held accountable for their actions, while also mitigating any guilt one might experience in pursuing punitive action. The fact that our punitive motivation manipulation had a stronger effect on our dehumanization measure than on our agency measure is somewhat consistent with such an account. Future research could investigate this possibility more thoroughly by manipulating the harmful agent's *capacity to understand the consequences of his or her actions*, independent of motivations to punish the target. Of course, it

is also possible that the null effect of punishment motivations on agency attributions is a true null effect, thus, research would certainly benefit from future attempts at replication.

One potential limitation of the current set of studies is that the vignettes we used as study materials may lack high levels of psychological realism and thus ecological validity. While the nature of our study materials offers considerable experimental control over confounding variables, future research should certainly go beyond short, hypothetical scenarios to examine the research questions we raise with methodologies that might allow for greater psychological realism.

One other avenue for future research is to examine whether the relationship between harmfulness and agency exhibits reverse causality. While the present research demonstrates that harmful individuals are ascribed less agency than benevolent and non-offending agents, it leaves open the possibility that highly agentive individuals are perceived as *less harmful* (the reverse causal direction). Preliminary support for this idea was found in the results of Study 5. Individuals rated the highly rational offender as less harmful (M = 6.16, SD = .69) compared to the less rational individual (M = 6.55, SD = .70), $F(1,111) = 9.25$, $p = .003$, $\eta^2_p = .077$. The inference here might be that rational offenders are more in control of their behavior (e.g., possess greater self-restraint) and therefore are less likely to cause inadvertent or incidental harmful consequences when executing their intentions. This would be consistent with the results observed that rational harmful agents are ascribed higher levels of agency (e.g., self-control) than less rational harmful agents. However, future research should test this hypothesis more systematically.

8. Conclusion

Thomas Kuhn wrote, "The proponents of competing paradigms practice their trades in different worlds." This research has attempted to unite the worlds of dehumanization and moral typecasting and provide a singular empirical test of these competing accounts regarding the conceptualized agency of harmful agents. The results demonstrated overwhelming support for a dehumanization account, and offered evidence contrary to assumptions made by moral typecasting theory. Rather than being typecasted as moral agents, harmful agents were attributed less agency than both non-offenders and benevolent agents. This occurs largely due to the inferences perceivers draw about the qualities possessed by someone who engages in cruel acts. Cruel agents are dehumanized and as a consequence are stripped of basic rights and protections generally afforded to all humans.

Appendix A. Key Measures from Studies 1-5

Harmfulness	Please rate the extent to which you perceive this person as having these qualities: Aggressive. Mean. Hostile. Peaceful (reverse coded). Gentle (reverse coded).
Moral Standing	How morally wrong do you think it would be for someone to harm this person? How morally wrong do you think it would be for someone to steal from this person? To what extent do you think this person deserves to be treated with compassion and fairness? To what extent do you think this person deserves to be protected from harm? If this person was endangered, how important would it be to protect this person?
Agency	This person appears to be capable of making plans and working toward goal. This person appears to be capable of trying to do the right thing and telling right from wrong. This person appears to be capable of remembering things. This person appears to be capable of understanding how others are feeling. This person appears to be capable of exercising self-restraint over desires, emotions or impulses. This person appears to be capable of thought. This person appears to be capable of conveying thoughts or feelings to others.
Patiency	This person appears to be capable of longing or hoping for things. This person appears to be capable of experiencing embarrassment. This person appears to be capable of feeling afraid or fearful. This person appears to be capable of feeling hungry. This person appears to be capable of experiencing joy. This person appears to be capable of experiencing physical or emotional pain. This person appears to be capable of having personality traits that make him unique from others. This person appears to be capable of experiencing physical or emotional pleasure. This person appears to be capable of experiencing pride. This person appears to be capable of experiencing violent or uncontrolled anger. This person appears to be capable of having experiences and being aware of things.
Denial of Human Nature	I feel like this person was open minded, like he could think clearly about things (reverse coded). I feel like this person was emotional, like he was responsive and warm (reverse coded). I feel like this person was superficial, like he had not depth. I feel like this person was mechanical and cold, like a robot.
Denial of Human Uniqueness	I feel like this person was refined and cultured (reverse coded). I feel like this person was rational and logical, like he was intelligent (reverse coded). I feel like this person lacked self-restraint, like an animal. I feel like this person was unsophisticated.
Rationality	To what extent would you say this person's actions are: Reasonable; Rational; Sensible; Understandable.
Motivation to Punish	I want to retaliate against this person for the incident. I want to take revenge on this person. I want to get even with this person.

Appendix B. Vignettes Used in Studies 2a-5

Study 2a

Harmful agent. John is a 29 year old man with brown hair and brown eyes. He works as a taxi driver in Miami. At night, when John is off work, he likes to spend his time collecting stray cats and dogs to torture and experiment on in his basement.

Benevolent agent. John is a 29 year old man with brown hair and brown eyes. He works as a taxi driver in Miami. At night, when John is off work, he likes to spend his time helping out with Meals on Wheels to deliver food to the needy.

Neutral agent. John is a 29 year old man with brown hair and brown eyes. He works as a taxi driver in Miami. At night, when John is off work, he likes to spend his time practicing guitar and writing music.

Study 2b

Harmful/likeable agent. David is a vegetarian. The reason he is a vegetarian is because he thinks humans are no more important than animals. David enjoys it when people who eat meat suffer. David sometimes puts extra-spicy hot sauce in people's hamburgers when they aren't looking. David's friends find him quite enjoyable to be around because he has a witty sense of humor and playful character.

Benevolent/unlikeable agent. David is a vegetarian. The reason he is a vegetarian is because he cares a great deal about animals and it upsets him that animals are killed for their meat when

alternative food products are available. David sometimes volunteers at the local animal shelter to help care for the animals. David's friends find him quite annoying to be around because he is always trying to get them to change the way they eat.

Neutral/likeable agent. David is a vegetarian. The reason he is a vegetarian is because he doesn't like the smell or taste of meat. David sometimes works at the outlet mall selling perfume and cologne. David's friends find him quite enjoyable to be around because he has a witty sense of humor and playful character.

Study 3

Highly harmful agent. Initrode is a cigarette and tobacco manufacturer. They operate primarily in Africa and parts of Southeast Asia. The mix of harsh tobacco and additional chemicals used in their products have been linked to a much higher incidence of lung-cancer and heart-disease than normal cigarette use.

Less harmful agent. Initrode is a cigarette and tobacco manufacturer. They operate primarily in Africa and parts of Southeast Asia. The mix of tobacco and chemicals used in their products have been linked to a much lower incidence of lung-cancer and heart-disease than normal cigarette use.

Moderately harmful agent. Initrode is a cigarette and tobacco manufacturer. They operate primarily in Africa and parts of Southeast Asia.

Study 4

Violent offender. Ryan Macey is a 41 year old Caucasian man from Sydney. On a Tuesday afternoon, he attempted to hijack a bus carrying 50 passengers by threatening the driver with a knife. Macey boarded the bus stopped in traffic, forced the driver off, and told passengers to stay put. He failed to start the bus before off-duty Senior Constable John Rider arrived. Macey lunged at the officer with a chisel before fleeing towards the Westfield shopping centre, where police lost track of him. Ryan was shirtless and seen on TV soon after the crime. He was arrested about 2.35pm when a resident phoned police saying they had been threatened by a man hiding out in their garage, and was found carrying a bum-bag containing a chisel, screwdriver, scissors and a 20 centimetre blade knife.

White-collar. Ryan Macey is a 41 year old Caucasian man from Sydney. He fleeced almost $127,000 from family and friends by stealing money they had given to him for investments. He stole this large amount of money from 6 clients between February 2002 and July 2004. He had encouraged clients, family and friends to invest in shares, managed investments and superannuation funds. Instead of investing the money, Macey transferred the funds into his own accounts and used the money for personal expenses. Ryan has been sentenced to 10 month s in jail before being released on a five year good behaviour bond.

Neutral. Ryan Macey is a 41 year old Caucasian man from Sydney. He works as a teacher at Charles Duncan Elementary School. He has worked there for the past 10 years and is currently teaching a grade 7 class. This upcoming Friday, his students will have a math test and so he has been focusing primarily on math this week. For the weekend he has plans to catch a play at the

local theater.

Study 5

High rationality/ Low motivation to punish. John is a 23 year old male. He is about six feet tall, with brown hair and blue eyes. John has taken to crime. He does this because he needs the money to feed himself. Imagine one day he mugs another gang-member, stealing everything he had and leaving him injured.

High rationality/ High motivation to punish. John is a 23 year old male. He is about six feet tall, with brown hair and blue eyes. John has taken to crime. He does this because he needs the money to feed himself. Imagine one day he mugs your mother, stealing everything she had and leaving her injured.

Low rationality/ Low motivation to punish. John is a 23 year old male. He is about six feet tall, with brown hair and blue eyes. John has taken to crime. He does this for fun. Imagine one day he mugs another gang-member, stealing everything he had and leaving him injured.

Low rationality/ High motivation to punish. John is a 23 year old male. He is about six feet tall, with brown hair and blue eyes. John has taken to crime. He does this for fun. Imagine one day he mugs your mother, stealing everything she had and leaving her injured.

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*, 347-356.

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology, 38*, 1202-1217.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*, 751-763.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126, 556-574.*

Arico, A. J. (2012). Breaking out of moral typecasting. *Review of Philosophy and Psychology, 3*, 425-438.

Ashworth, A. (2010). *Sentencing and criminal justice*. New York: Cambridge University Press.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*, 193-209.

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

Bastian, B., & Haslam, N. (2010). Excluded from humanity: The dehumanizing effects of social ostracism. *Journal of Experimental Social Psychology, 46*, 107-113.

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology, 50*, 469-483.

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. M. (2012). Don't mind meat? The

denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin, 38,* 247-256.

Baumard, N. (2011). Punishment is not a group adaptation. *Mind & Society, 10,* 1-26.

Baumard, N., Andre, J., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36,* 59-78.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83,* 284-299.

Castano, E., & Giner-Sorolla, R. (2006). Not quite human: Infrahumanization in response to collective responsibility for intergroup killing. *Journal of Personality and Social Psychology, 90,* 804-818.

Christopher, P. P., & Pinals, D. A. (2010). Capacity to consent to sexual acts: Understanding the nature of sexual conduct. *Journal of the American Academy of Psychiatry and the Law, 38,* 417-420.

Cohen, R. J. (2014). Brand personification: Introduction and overview. *Psychology and Marketing, 31,* 1-30.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4,* 308-315.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108,* 281-289.

Darley, J. M., Klosson, E. C., & Zanna, M. P. (1978). Intentions and their contexts in the moral judgments of children and adults. *Child Development, 49,* 66-74.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice.

Personality and Social Psychology Review, *7*, 324-336.

Epley, N., & Waytz, A. (2009). Mind perception. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.). *Handbook of Social Psychology* (5th ed. Vol. 1, pp. 498-541). New York: Wiley.

Fitzgerald, P. J. (1962). *Criminal law and punishment*. Oxford: Clarendon Press.

Gert, B. (2004). *Common morality: Deciding what to do*. New York: Oxford University Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029-1046.

Gray, K. (2010). Moral transformation: Good and evil turn the weak into the mighty. *Social Psychological and Personality Science*, *1*, 253-258.

Gray, K. (2014). Harm concerns predict moral judgments of suicide: Comment on Rottman, Kelemen and Young (2014). *Cognition, 133*, 329-331.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology*, *3*, 405-423.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23*, 206-215.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*, 505-520.

Gray, K., & Wegner, D. M. (2011). To escape blame, don't be a hero-Be a victim. *Journal of Experimental Social Psychology, 47*, 516-519.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the

uncanny valley. *Cognition, 125*, 125-130.

Greene, J. D. (2012). *The moral brain and how to use it*. New York: Penguin Group.

Hart, H. L. A. (1968). *Punishment and responsibility: Essays in the philosophy of law*. New York: Oxford University Press.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*, 252-264.

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition, 26*, 248-258.

Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology, 65*, 399-423.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*, 1767-1770.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.

Kervyn, N., Fiske, S. T., & Malone, C. (2012). Brands as intentional agents framework: How perceived intentions and ability can map brand perception. *Journal of Consumer Psychology, 22*, 166-76.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*, 309-324.

Kozak, M. N., Marsh, A. A., & Wegner, D. M. (2006). What do I think you're doing? Action

identification and mind attribution. *Journal of Personality and Social Psychology, 90*, 543-555.

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*, 181-192.

Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup glorification, moral disengagement, and justice in the context of collective violence. *Personality and Social Psychology Bulletin, 36*, 1115-1129.

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

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147-186.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101-121.

Piazza, J., Landy, J. F., & Goodwin, G. P. (2014). Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition, 131*, 108-124.

Pinker, S. (2012). *The better angels of our nature: Why violence has declined*. New York: Penguin Books.

Robinson, P. (1984). *Criminal law defences* (Vols 1 & 2). St Paul: West Publishing Company.

Singer, P. (2011). *Practical ethics*. (3rd ed.). New York: Cambridge University Press.

Sinnott-Armstrong, W. (2009). *Morality without God?* New York: Oxford University Press.

Sousa, P., Holbrook, C., & Piazza, J. (2009). The morality of harm. *Cognition, 113*, 80-92.

Sousa, P., & Piazza, J. (2014). Harmful transgressions qua moral transgressions: A deflationary view. *Thinking & Reasoning, 20,* 99-128.

Sytsma, J., & Machery, E. (2012). The two sources of moral standing. *Review of Philosophy and Psychology, 3,* 303-324.

Thomson, M., MacInnis, D. J., & Park, C. W. (2005). The ties that bind: Measuring the strength of consumers' emotional attachments to brands. *Journal of Consumer Psychology, 15,* 77-91.

Turiel, E. (1983). *The development of social knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

Vasquez, E. A., Loughnan, S., Gootjes-Dreesbach, E., & Weger, U. (2014). The animal in you: Animalistic descriptions of a violent crime increase punishment of perpetrator. *Aggressive Behavior, 40,* 337-344.

Viki, G. T., Fullerton, I., Raggett, H., Tait, F., & Wiltshire, S. (2012). The role of dehumanization in attitudes toward the social exclusion and rehabilitation of sex offenders. *Journal of Applied Social Psychology, 42,* 2349-2367.

Waytz, A., & Epley, N. (2012). Social connection enables dehumanization. *Journal of Experimental Social Psychology,* 48, 70-76.

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

Footnotes

[1]For example, Gray and Wegner (2009) write, "The perception of humans and other entities along distinct dimensions of moral agency and moral patiency has been observed by Gray, Gray, and Wegner (2007). In this factor analytic study, the authors explored the dimensions of mind perception. (…) Participants compared pairs of entities on each of 18 mental qualities (e.g., the ability to feel hunger), and analyses of mean judgments revealed a two-dimensional solution corresponding in key aspects to the constructs of moral agency and moral patiency. A dimension termed Experience included many mental qualities indicating moral patiency: the abilities to feel hunger, fear, pain, pleasure, rage, and desire; to have personality and consciousness; and to feel pride, embarrassment, and joy. A dimension termed Agency included characteristics more relevant to moral agency: abilities to have self-control, morality, memory, emotion recognition, planning, communication, and thought" (p. 506).

[2] Gray and Wegner (2011, p. 518) write, "Previous moral agents, *whether they did good or evil,* remain typecast as agents for future misdeeds and are punished accordingly" (italics added). In discussing the results of studies testing the moral typecasting effects of good and bad deeds, Gray (2010) writes, "In Experiment 1, individuals who did good possessed more agency. Experiment 2 found that those who imagined themselves doing good or evil were more agentic than those who imagined themselves doing something neutral" (p. 257).

[3] It is also clear from the methodology of Gray and Wegner (2009), which included both harmful and benevolent moral agents (see Studies 2, 3b, 3c, and 7), that Gray and Wegner understand the moral typecasting hypothesis to apply similarly to harmful and benevolent agents (see esp. p. 518) "…we looked at responses to a range of good and evil actors, again finding no

noteworthy discontinuities in the occurrence of moral typecasting across a range of good and bad agents.".

[4] The authors acknowledge the comments of two anonymous reviewers as to this possible alternative explanation underlying the attribution of agency.

[5] While agent condition remained a significant predictor, likeability was also a significant predictor of cognitive agency attributions, $\beta = .66$, $t(93) = 7.69$, $p < .001$, but not of activity agency attributions, $\beta = -.09$, $t(93) = -.69$, $p = .49$. Likeability also predicted judgments of the agent's moral standing, $\beta = .86$, $t(93) = 8.65$, $p < .001$.

[6] There was a significant main effect of agent condition on moral standing ratings when likeability was not used as a covariate, $F(2,93) = 19.59$, $p < .001$, $\eta^2_p = .296$. John was attributed lower levels of moral standing in the harmful agent condition ($M = 3.68$, SD = 1.96) than in the neutral ($M = 5.95$, SD = 1.33), $p < .001$, and benevolent agent conditions ($M = 5.88$, SD = 1.65), $p < .001$, whereas John was attributed similar levels of moral standing in the neutral and benevolent agent conditions, $p = .86$.

[7] The indirect effect of agent condition on judgments of the agent's moral standing via attributions of agency still holds using a single collapsed measure of agency that includes both mind perception and activity traits, indirect effect = -.72 (95% bias corrected CIs [-1.26, -.30]).

Supplementary Materials

Study 4 Supplements

Factor analysis results from Study 4 including agency and denial of humanness items.

We submitted the agency and denial of humanness scales to a principal components factor analysis with Varimax rotation. The analysis produced a two-factor solution, explaining 69.77% of the total variance, with the eigenvalues for the two factors equal to 8.92 and 1.54. The first factor (eigenvalue = 8.92) contained most of the denial of humanness items (6 out of 8) and three of the agency items ("capable of trying to do the right thing and telling right from wrong"; "capable of understanding how others are feeling"; "capable of exercising self-restraint over desires, emotions or impulses"). The second factor (eigenvalue = 1.54) contained three agency items ("capable of making plans and working toward a goal"; "capable of thought"; "capable of remembering things") and one reverse-coded denial of humanness item ("the person in the story was rational and logical, like he was intelligent"). One dehumanization item ("the person in the story was unsophisticated") and one agency item ("capable of conveying thoughts or feelings to others") cross-loaded on the two factors.

Study 5 Supplements

Factor analysis results from Study 5 including agency and patiency items.

We submitted the agency and patiency scales to a principal components factor analysis with Varimax rotation, much like in Gray et al.'s original study, using parallel analysis as our extraction method. We compared the eigenvalues generated by the parallel analysis with those generated by the principal components analysis. This comparison suggested a two-factor solution, explaining 58.7% of the total variance. The first factor (eigenvalue = 7.72) contained all of the agency items. The second factor (eigenvalue = 2.27) contained all but one of the patiency

items; the item "experiencing violent or uncontrolled anger" failed to load on either factor, but we included it in our index of patiency since including it did not reduce the overall reliability of the scale. One patiency item ("capable of longing or hoping for things") cross-loaded with the agency factor.

Factor analysis results from Study 5 including agency and denial of humanness items.

We submitted the agency and denial of humanness scales to a principal components factor analysis with Varimax rotation. The analysis produced a three-factor solution, explaining 67.21% of the total variance, with the eigenvalues for the three factors equal to 6.27, 2.62 and 1.19. The first factor (eigenvalue = 6.27) contained all but one of the agency items; the item "capable of exercising self-restraint over desires, emotions or impulses" cross-loaded on the second factor. The second factor (eigenvalue = 2.62) contained three reverse-coded denial of humanness items ("this person was emotional, like he was responsive and warm"; "this person was refined and cultured"; "this person was rational and logical, like he was intelligent"). The third factor (eigenvalue = 1.19) contained three other denial of humanness items ("this person was superficial, like he had not depth"; "this person was mechanical and cold, like a robot"; "this person was unsophisticated"). One dehumanization item ("this person lacked self-restraint, like an animal") and one agency item ("capable of exercising self-restraint over desires, emotions or impulses") cross-loaded on the two factors.

Results for patiency from Study 5.

There was a significant main effect of rationality condition on patiency ratings, $F(1,107) = 4.05$, $p = .047$, $\eta^2_p = .036$, that was consistent with the results for agency. The less rational agent was attributed lower levels of patiency ($M = 4.81$, SD = 1.20) than the more rational agent ($M = 5.23$, SD = .96). However, there was no significant main effect of motivations to punish on

patiency ratings, $F(1,107) = .80$, $p = .37$, $\eta^2_p = .007$. Additionally, the rationality and motivation manipulations did not interact to affect patiency, $F(1,105) = 0.00$, $p = .99$, $\eta^2_p = .000$.

We ran mediation analyses including patiency as a simultaneous mediator, in addition to agency and denial of humanness, of the effects of rationality and, separately, motivations to punish, on moral standing judgments. Multicollinearity between patiency and agency (tolerance = .619, VIF = 1.62) and patiency and denial of humanness (tolerance = .935, VIF = 1.07) did not appear to be an issue. The indirect effect of offense rationality on judgments of moral standing operated exclusively through denial of humanness attributions, $ab_{dehumanization} = -.22$ (95% bias corrected CIs [-.53, -.05]), and not through agency, $ab_{agency} = .02$ (95% bias corrected CIs [-.21, .26]) or patiency, $ab_{patiency} = -.15$ (95% bias corrected CIs [-.51, .01]). Similarly, the indirect effect of motivation to punish condition on judgments of moral standing operated only through denial of humanness attributions, $ab_{dehumanization} = .19$ (95% bias corrected CIs [.03, .47]), and not through agency, $ab_{agency} = .01$ (95% bias corrected CIs [-.07, .18]) or patiency, $ab_{patiency} = .08$ (95% bias corrected CIs [-.04, .35]).

Main analysis from Study 5 using the overall mind perception index.

There was a significant main effect of rationality condition on the combined agency/patiency index, $F(1,110) = 5.80$, $p = .018$, $\eta^2_p = .050$. The less rational agent was attributed lower levels of mind ($M = 4.20$, SD = 1.17) than the more rational agent ($M = 4.67$, SD = .93). However, there was no significant main effect of motivation to punish on the combined mind perception index, $F(1,110) = .47$, $p = .47$, $\eta^2_p = .004$. Additionally, there was no interaction effect of rationality and motivation to punish on the combined mind perception index, $F(1,110) = .50$, $p = .48$, $\eta^2_p = .005$.