Interconnections Between Perceptions of Blame, Mind, and Moral Abilities

Alex Ropes
Macalester College, aropes@macalester.edu

Steve Guglielmo
Macalester College, sgugliel@macalester.edu

Follow this and additional works at: http://digitalcommons.macalester.edu/psychology_honors

Part of the Applied Behavior Analysis Commons, Social Psychology Commons, and the Theory and Philosophy Commons

Recommended Citation
http://digitalcommons.macalester.edu/psychology_honors/36

This Honors Project is brought to you for free and open access by the Psychology Department at DigitalCommons@Macalester College. It has been accepted for inclusion in Psychology Honors Projects by an authorized administrator of DigitalCommons@Macalester College. For more information, please contact scholarpub@macalester.edu.
Interconnections between perceptions of blame, mind, and moral abilities

Macalester College

Alex Ropes
Abstract

Theories of blame, mind, and moral attribution consider an individual’s perceived agency, operationalized in part as perceived intentionality and self-control. People with autism spectrum disorder (ASD) may display social deficits and a greater tendency to engage in problem behavior (PB; American Psychiatric Association, 2013) than neurotypical (NT) people, which may lead people to perceive that individuals with ASD act less agentically. Study 1 shows that the mitigated perceived agency of people with ASD leads to mitigated blame attribution. In addition to perceived agency, theories of mind and moral attribution account for perceptions of an individual’s capacity to experience emotions, pleasure, and suffering. Based upon these forms of perception, Gray et al.’s (2007) theory of mind perception (TMP) states that minds are perceived along the dimensions of agency and experience. Similarly, Gray, Young, and Waytz’s (2012) theory of dyadic morality (TDM) states that a person’s moral status is perceived along the dimensions of moral agency and patiency. While these dimensional pairs are highly similar, the TMP states that its proposed dimensions are independent of each other while the TDM states that its proposed dimensions are inversely related. Studies 2 and 3 generated support for the prediction that these dimensions are independently related, as proposed by TMP, while the inverse relationship posited by the TDM did not receive support.
Interconnections between perceptions of blame, mind, and moral abilities

Blame is a multi-faceted social phenomenon, used to set and affirm norms, and to evaluate events and agents (Malle, Guglielmo, & Monroe, 2014). As a moral judgment, the amount of blame attributed to individuals can have profound effects on their lives—for example, higher blame attribution leads to the assignment of more severe punishments (Carlsmith, Darley, & Robinson, 2002). As a result, understanding the ways in which people attribute blame to others, and ensuring that the appropriate amount of blame is assigned, is integral to ensuring the fair and equitable treatment of individuals. Malle et al. (2014) have proposed a Path Model of Blame outlining the blame attribution process, and this model provides a thorough consideration of the various sub-components that influence blame attribution. Included in this model are considerations of an agent’s causality, intentionality, obligation to have acted otherwise (hereafter referred to as “obligation”), capacity to have acted otherwise (hereafter referred to as “capacity”), and reasons for acting. These subcomponents are sequenced in this model, reflecting the intuitive process by which individuals assign blame: if an agent is determined to have acted intentionally, their reasons for acting are considered; if they were determined to have acted unintentionally, their obligation and capacity are considered.

These judgments are made uniquely with regard to each novel norm-violation. For example, when an individual accidentally kills another person, their obligation and capacity are judged much differently than if they had accidentally knocked over a lamp. Yet, certain factors or identities may exert stable cross-situational effects on components of a path model. For example, children are typically assigned less blame for norm violations (e.g., making a hurtful comment) than adults. This differential blame attribution arises from a number of differences between children and adults. First, it is recognized that an understanding of norms is not innate,
and is rather learned by individuals as they develop (Jensen, Vaish, & Schmidt, 2014). Furthermore, individuals’ brains, and cognitive capacities, do not finish developing until their third decade of life (Gogtay et al., 2004), leading children to possess lower levels of self-control (Arain et al., 2013) and foresight (i.e., the ability to consider the long-term consequences of actions; Lewis, 1981) than adults. Due to these factors, most individuals would perceive that children act less intentionally than adults, and do not have the same obligation or capacity to avoid committing norm-violating actions as adults do. This recognition leads to mitigated blame attribution, which has in turn been codified into our legal system, such that children are typically sentenced much less harshly than adults an identical crime (American Bar Association, 2007).

An individual’s age is only one factor that may exert such stable effects on blame judgments, though, and to ensure the equitable treatment of all people (e.g., in the context of criminal sentencing), one must recognize other such factors. One such factor is autism spectrum disorder (ASD), the characteristics of which may lead individuals to differentially attribute blame to a person with ASD, as compared to a neurotypical (NT) individual, when a norm violation was putatively related to ASD. Before addressing this possibility, though, I will further explain the path model in order to establish the theoretical framework in which I will be working. Then, I will address which aspects of ASD may affect the blame attribution process. Finally, I will consider how ASD may also affect the dehumanization process, which also takes into account many of the factors that drive the blame attribution process.

**Blame**

As mentioned, Malle et al.’s (2014) path model of blame provides a unified framework that sequences various factors that affect the blame attribution process, and a more robust understanding of this model is necessary in order to identify factors that may exert stable effects
on blame attribution. According to the path model, the blame attribution process begins when an individual detects a norm violation and causally links it to an agent. Once it has been confirmed that an agent caused a norm violation, their intentionality is then considered. If the individual is perceived to have acted intentionally, their reasons for acting are taken into account. For example, if an individual acted for an asocial, vengeful, or selfish reason (Reeder, Kumar, Hesson-McInnis & Trafimow, 2002) or if their action predicts further norm violations (Tetlock et al., 2007) blame judgments are exacerbated. If an individual acts in self-defense (Finkel, Maloney, Valbuena, & Groscup, 1995) or for the greater good (Lewis et al., 2012), blame judgments are typically mitigated. In this vein, if a woman shoves a man to the ground in order to steal his money and run away (i.e., an asocial and selfish reason), she will be blamed more than if she shoved the man to the ground because he was trying to harm her (i.e., she acted in self-defense).

If the agent acted unintentionally, their obligation to have acted otherwise is first considered. For example, the higher an individual is in a social hierarchy, the stronger their perceived obligation is for preventing negative outcomes (Hamilton, 1986) and the more they are blamed for unintentional norm violations (Gibson & Schroeder, 2003). If an agent was not perceived to have had an obligation to prevent the norm violation, they are assigned little to no blame. Conversely, if they are determined to have had such an obligation, their capacity to have prevented the norm violation is considered. This functions such that if the agent had the capacity to have prevented the outcome, or possessed the foresight to realize that the negative event would occur, they are assigned more blame (e.g., when negligent repairs lead to a car accident, the auto-mechanic is blamed; Lagnado & Channon, 2008). Overall, blame mitigation typically occurs if an intentional action was committed for acceptable reasons (e.g., a terrorist was killed
to save thousands), if the action was unintentional (e.g., an individual accidentally trips another person), or an individual had a mitigated capacity and/or obligation to have acted other than how they did (e.g., a man is hired to clean out an attic, and throws out a tattered tapestry, not realizing that it is a treasured family heirloom).

**Autism Spectrum Disorder**

ASD is characterized by a variety of social deficits, restricted or repetitive behaviors and interests, and a sensitivity to change and certain stimuli (American Psychiatric Association, 2013). Although the social and behavioral deficits related to ASD map onto the Path Model in unique ways (such that people with ASD may be perceived as acting with less capacity and intentionality than neurotypical individuals due to these deficits), the extant literature has not yet addressed the possibility that ASD may exert stable effects on blame judgments. Although this particular link has not yet been explored, blame and punishment are positively correlated (Carlsmith et al., 2003), and current behavioral management programs advocate for mitigated punishments towards people with ASD (Carr et al., 2002), suggesting that blame attribution may also be mitigated. Furthermore, theories of blame and punishment share similar frameworks, such that both take into account an actor’s intentionality, capacity, and reasons for acting (Carlsmith, Darley, & Robinson, 2002; Darley & Pittman, 2003; Kane, Joseph, & Tedeschi, 1977, Lagnado & Channon, 2008; Malle et al., 2014). Consequently, a consideration of how these factors drive mitigated punishment may in turn allow one to better hypothesize about how these factors may affect blame attribution.

The literature most often addresses how people with ASD are punished in response to displays of problem behavior (PB), a broad term that refers to any disruptive behavior—such as self-injury, aggression, or tantrums—exhibited by an individual (Hagopian, 2007). Sixty-four to
93% of children with ASD display PB (Kozlowski, Sipes, & Matson, 2012), and numerous behavioral interventions have been developed to manage and reduce rates of PB. Positive behavioral supports (PBS) are one of the current leading behavioral management programs used for individuals with developmental and intellectual disabilities. PBS were developed to minimize displays of PB and maximize the quality of life of people with developmental and intellectual disabilities, and they have been implemented with success among populations with ASD (Carr et al., 2002; Safran & Oswald, 2003). Although PBS can be implemented in any environment, it has been most widely used in the public school system as a result of the implementation of the Individuals with Disabilities Education Improvement (IDEA) Act in 2004, which legally mandated that schools have systems in place to accommodate the specific needs of individuals with disabilities. As a result, the majority of the literature on PBS addresses how it functions in a school context.

PBS explicitly differs from traditional disciplinary responses to PB. Traditional forms of discipline used with NT individuals are typically reactionary and centered on negative reinforcement (e.g., time-outs for children or speeding tickets for adults). Individuals are expected to understand social norms and legal rules, and if one of these rules is broken, an appropriate punishment is meted out (Darley & Pittman, 2003). In contrast to this, PBS takes a preventative stance, tailoring environments to help individuals to avoid situations in which they may commit norm violations. The focus of PBS interventions is “fixing problem behavior contexts, not problem behavior” (Carr et al., 2002, p. 8). For example, if a music class is completing a unit on percussion instruments, a student with an extreme sensitivity to loud noises may be allowed to take periodic sensory breaks during class, or complete an alternative
educational activity in a different classroom. In this way, the environment is tailored so PB does not have the chance to arise.

The discipline used by PBS programs can be further considered in relation to two predominant theories explaining how individuals punish others: just deserts and deterrence theory (Carlsmith et al., 2002). According to just deserts (or retributive) punishment, the punishment assigned to an individual should be equivalent in extremity to the norm violation committed; according to deterrence theory, the punishment assigned an individual should be designed to prevent future similar norm violations. For example, consider a pharmaceutical executive who illegally overcharges cancer patients for their treatment and makes millions of dollars for himself. If an individual were guided by just deserts motives, they would assign this executive a punishment deemed to be equivalent to the monetary and emotional harm he caused the victims of his crime. If an individual were guided by deterrence motives, their assigned punishment would not necessarily be equivalent to the harm caused, but it would be highly publicized, to deter others from committing a similar wrong in the future.

Although many individuals cite deterrence as their principal motive when assigning punishment, most punishments are in fact informed by a just deserts motive (Carlsmith et al., 2002; Darley & Pittman, 2003). This effect is qualified by the intentionality of the action, though, such that while the majority of intentional wrongdoings are responded to with just deserts punishment, accidental wrongdoing is responded to with utilitarian, deterrence-oriented punishment. Interestingly, the disciplinary framework used by PBS reflects deterrence-oriented principles. While PBS primarily advocates for tailoring environments to prevent the display of PB in the first place, when PB does arise, deterrence-oriented disciplinary methods are invoked (Carr et al., 2002). If an NT student has an outburst during class they would likely be assigned a
time-out of equivalent length to the severity of the outburst, reflecting just deserts punishment. If a student with a disability in a PBS program were to have an outburst, their punishment would be designed to optimally reduce the likelihood of such behavior in the future, reflecting deterrence-oriented punishment. For example, the student might lose access to a preferred activity (e.g., computer time) that will best incentivize them to avoid such outbursts in the future. In this case scenario, the severity of the punishment may not be equivalent to the severity of the outburst, but it will be designed to best prevent the display of such an outburst in the future.

Despite the implementation of deterrence-oriented principals in PBS programs, the literature addressing PBS is concerned with its practical application, rather than the theories driving it. Rather than taking up the philosophical question of whether people with disabilities deserve mitigated punishments, or the psychological question of what factors drive these mitigated punishments, the literature on PBS is focused on the utilitarian outcome that its methods improve the quality of life for both people with disabilities as well as the people around them through the reduction of PB (Carr et al., 2002; Safran & Oswald, 2003). Consequently, the extant literature has not addressed whether the disciplinary model of PBS is founded on the presumption that individuals falling within the scope of the program act less intentionally than NT individuals, even though its focus on deterrence-oriented punishment suggests that this may be the case (Darley & Pittman, 2003). Relatedly, the literature also has not yet addressed whether the mitigated punishments advocated by the program are driven by, or associated with, mitigated blame attribution, despite salient links between punishment and blame attribution. A consideration of how the aspects of ASD may map onto Malle et al.’s (2014) path model allows for a more rigorous theoretical consideration of the possibilities raised here.
Behavioral dispositions. ASD may affect the behavioral disposition of individuals in a number of ways, such as through the manifestation of restricted and repetitive behaviors (e.g., refusing to eat foods of a certain texture) and hypersensitivities to certain stimuli (e.g., refusing to wear any type of pant other than sweatpants; American Psychiatric Association, 2013; Wilson et al., 2013). Disruptions to schedules or rituals, or exposure to a hyper-sensitized stimulus, may lead an individual with ASD to engage in problem behavior. For example, a small, but sudden, shift to the daily schedule of a middle school student (such as missing a class for a school-wide assembly) could lead the student to display PB. While such behavior would certainly be deemed blameworthy if it were engaged in by an NT student (it is expected that students follow the school’s instructions), would it be considered equivalently blameworthy when an individual with ASD engages in it?

To begin, one must determine whether this action was intentional or unintentional. Malle and Knobe (1997) outline five factors that comprise an intentional action: an agent’s desire, belief, intention, awareness, and skill. Desire refers to the agent consciously hoping to attain a certain outcome and belief refers to their knowledge of the consequences that are linked to executing an action. An agent’s “intention” refers to an agent enacting their desire and belief by executing an action to attain a goal. For example, a basketball player may desire to score a free throw, believe that if she throws the ball then it will move in the direction that she throws it in, and, based upon this knowledge, intend to throw the ball. If the basketball player does indeed score a free throw, two final aspects must be taken into account before her intentionality is determined: her awareness and skill. Awareness refers to an individual’s self-awareness while executing an action while skill refers to their actual ability to accomplish their desired goal. For example, if a seventh grade, JV basketball player scores a half-court, buzzer shot, most
spectators would discount how intentionally she acted, recognizing that she does not possess the skill to accomplish such a feat, and that she rather got lucky.

Having addressed the five aspects of intentionality, one can now determine whether an individual with ASD acts intentionally or unintentionally when displaying PB influenced by aspects of ASD. As mentioned, people with ASD have a greater tendency to display PB than NT individuals due to deficits stemming from ASD (Kozlowski et al. 2012). In other words, people with ASD may have a lower threshold at which they display PB than NT individuals. Furthermore, the majority of PB displayed by individuals with ASD has an identifiable, immediate cause, as evidenced by the literature’s focus on identifying antecedent and reinforcing events, and stimuli that may trigger the display of PB (Matson & Nebel-Schwam, 2007). That is, the PB engaged in by people with ASD is not premeditated (e.g., forming a plan to destroy someone’s property), but rather arises in the heat of the moment (e.g., tantruming in response to a sensitized stimulus). In these cases, the question arises of whether PB was engaged in impulsively and with mitigated intentionality, or if it was fully intentional.

Consider the case of a student with ASD who shoves his teacher after becoming angered by a change to his daily schedule. In this case, the student clearly possessed the skill to shove his teacher, and understood (i.e., “believed,” to use Malle & Knobe’s (1997) terminology) that by thrusting his arms forward his teacher would be pushed back. Presumably, the student was also consciously aware of what he was doing and intended to shove the teacher. Lastly, then, one must consider the child’s desire. When a person commits a norm violating action, their desire to do so must override their self-regulatory capacities. For example, if a young girl is taunted by another child, she may be tempted to hit this child, but she does not because she knows that it is wrong to hit another person. But if this antagonist continues to taunt her day after day, her desire
to hit the child may continue to increase until it overrides her desire not to break the rules, and she finally hits the other child. Among individuals with ASD, this component of intentionality may be affected, such that individuals with ASD have a higher proclivity towards exhibiting PB in response to certain stimuli, and thus the threshold their desire must reach to engage in that action is much lower than the threshold of an NT person. According to Malle and Knobe (1997) mitigated desire should in turn mitigate perceived intentionality. Furthermore, this proclivity towards displaying PB also maps onto the capacity component of the path model, such that it reduces the capacity that individuals with ASD have to prevent the display of PB. According to Malle et al.’s (2014) path model, if people indeed perceive that a person with ASD acts with less intentionality and capacity than an NT person, they should also attribute less blame to the person with ASD.

**Social deficits.** Social deficits related to ASD can be grouped into two categories: theory of mind (ToM) related deficits and general social deficits. With regard to the first category, ToM is an individual’s ability to recognize, infer, and make sense of mental states, both their own and those of conspecifics (Premack & Woodruff, 1978). ToM perceptions and judgments pervade social interactions, affecting everything from the way that people attend to movement (Teufel, Fletcher, & Davis, 2010) to how they respond in a socially appropriate manner to others’ behaviors (Sigman, Kasari, Kwon, & Yirmiya, 1992). Classically, a false-belief test (in which a participant recognizes that another individual may hold an incorrect, subjective belief that is different from their own, objectively correct belief) has been used as the gold standard to test for ToM (Baron-Cohen, 1985). Although the false-belief test is still used to identify advanced ToM abilities, researchers now realize that ToM is not comprised solely of this ability, and a larger battery of tests are now administered to assess individuals’ ToM capabilities. While most
children pass these tests by age five, children with ASD do not consistently pass ToM tests until age 13 (Baron-Cohen, 1985; Happé, 1995; Peterson, Wellman, & Liu, 2004).

As opposed to the narrow conception of ToM originally proposed by Premack and Woodruff (1978) and tested by Baron-Cohen (1985), Wellman and Liu (2004) have proposed that ToM is composed of five distinct aspects, which are developed chronologically (although, interestingly, the last two capacities are achieved in reverse order for individuals with ASD). These are an individual’s ability to understand: “a) diverse desires, b) diverse beliefs, c) perceptual access to knowledge, d) false belief, and e) hidden emotion” (Peterson, Wellman, & Liu, 2005, p. 504). While a dissection of each of these components of ToM is not necessary within the scope of this paper, outlining these individual aspects of ToM illustrates its robust nature and the variety of ways by which related deficits may lead to norm violations. For example, recognizing that others can have desires and beliefs different than our own is necessary in order to avoid being perceived as selfish (e.g., when planning a vacation, a person asks for and considers the opinions of all of their friends when choosing a destination), and recognizing hidden emotion allows us to avoid being socially maladroit (e.g., realizing that a person is upset even when they say they are okay).

Many individuals with ASD are eventually able to pass the gamut of ToM tests (albeit at a later age than NT individuals) but several researchers have proposed that they accomplish this through the creation of workarounds (i.e., “social algorithms”) that compensate for their inability to utilize intrinsic ToM (Baez et al., 2012; Happé, 1995; Peterson et al., 2004). If these workarounds fully compensate for ToM deficits, then individuals with ASD would possess the equivalent capacity to navigate social situations as NT people, likely also leading to the
equivalent attribution of blame for social norm violations. Thus, a further consideration of the
efficacy of these workarounds is necessary.

Happé (1995) was one of the first psychologists to propose this workaround hypothesis. When studying the age at which children with ASD and NT children pass ToM tasks, he found that not only are children with ASD significantly older than NT children when they are able to pass equivalent tasks, but that they also possess a significantly higher verbal mental age (VMA) – a standardized measure of intelligence – when they pass these tasks. Furthermore, all participants with ASD passed both ToM tests administered in the study if they had a VMA of six years and nine months or above. Based on these findings, Happé (1995) hypothesized that participants with ASD were solving the tasks in a “verbally mediated fashion” (p. 852). That is, rather than spontaneously and subconsciously evoking ToM like NT participants, participants with ASD were consciously, verbally working through ToM tasks, solving them like a puzzle. Sigman et al. (1992) obtained similar results, and formed a similar hypothesis, after finding that children with ASD required significantly greater cognitive abilities than their NT counterparts in order to pass ToM tests, again suggesting that they may have been creating algorithms that were used to interpret social situations.

Yet, these workarounds may not make up for the lack of intrinsic ToM abilities. Other researchers have found that although individuals with ASD may be able to pass ToM tests in a laboratory setting, they may still be unable to spontaneously (Senju, 2012) or consistently (Scheeren, Rosny, Koot, & Begeer, 2013) compensate for ToM abilities in organic social settings. Some of the stronger evidence demonstrating disparities between the spontaneous activation of ToM comes from experiments using eye-tracking technology. Southgate, Senju, and Csibra (2007) had two-year old infants watch a video depicting a false belief task. They found
that when the actor returned to collect their object, infants showed anticipatory looking at the
location the actor had left their object at, even though the infants knew that the object had since
been moved to a new location. When Senju (2012) replicated this study with adults with ASD,
they found that they showed significantly less anticipatory looking than an NT control group.
This supports the hypothesis that although individuals with ASD may be able to pass formal
ToM tests, this may be accomplished through the creation of workarounds that do not fully
compensate for ToM deficits in organic, daily situations.

Moving past ToM deficits, several other researchers have also found that even when
individuals with ASD pass ToM tests, they still display numerous more general social deficits.
Klin et al. (2002) found that, while controlling for ToM abilities, individuals with ASD still
demonstrate a decreased capacity to read facial affect, recognize and make sense of implicit
social cues (e.g., irony and sarcasm), and read non-verbal social cues (e.g., pointing to an object
of interest) as compared to NT individuals. Additionally, although empathy is often understood
to be derivative of ToM, Peterson (2014) found that even when participants passed ToM tests,
they were still rated as significantly less empathetic than their NT conspecifics. This latter
finding is of particular relevance, as empathy is understood to play a fundamental role the
development and maintenance of social norms (Jensen, Vaish, & Schmidt, 2014). Thus, the
deficits to ToM abilities coupled with the general social deficits displayed by people with ASD
have the potential to considerably affect the capacity of an individual with ASD to foresee a
potential social norm violation. If an individual is unable to read another’s affect, observe their
implicit social cues, or hypothesize about their mental state, they may miss out on important
contextual cues that guide appropriate social interactions.
In sum, the extant literature has identified numerous social deficits, and deficits to ToM, that people with ASD may display. Although the latter of these may be counteracted in laboratory settings through the development of workarounds, they may still affect a person with ASD’s functioning in organic social settings. As with the display of PB, these deficits again map onto both the intentionality and capacity components of the path model, such that they may lead a person to unintentionally break a social norm or offend another person, and reduce their capacity to recognize that what they say or do may be perceived as offensive. If participants perceive that individuals with ASD act with less intentionality or capacity than an NT person due to the social deficits, and ToM deficits, related to ASD, according to the path model they will in turn assign mitigated blame.

**Dehumanization**

In the same way that blame judgments are influenced by perceptions an individual’s cognitive capabilities –to the extent that these capabilities influence portions of the path model, such as the agent’s intentionality or capacity– so too is dehumanization. Consequently, current theoretical models of blame (Malle et al., 2014) and dehumanization (Haslam, 2006) share several theoretical similarities. Dehumanization has been conceptualized in a variety of ways throughout the years, but all theories agree that dehumanization occurs when an individual is seen as less human than others. Initially, dehumanization was only considered to arise in egregious circumstances (Haslam, 2006), such as the treatment of Jews during the Nazi regime. But since the turn of the millennium, a number of other, subtler, conceptualizations of the phenomenon have been developed, with the aforementioned form of dehumanization now recognized as blatant dehumanization (Kteily, Bruneau, Waytz, & Cotterill, 2015).
Leyens et al. (2001) first proposed a theory of implicit dehumanization with their concept of infrahumanization. They define this as a “process by which people consider their ingroup as fully human and outgroups as less human and more animal-like” (Leyens, Demoulin, Vaes, Gaunt, & Paladino, 2007, p. 140). Leyens et al. (2001) propose that infrahumanization occurs in the way that individuals attribute emotions to others, and within their theory, they differentiate between primary and secondary, or human uniqueness (HU), emotions. Primary emotions can be attributed to both humans and animals (for example, a person may describe their pet dog as sad or happy) whereas HU emotions can only be attributed to humans (for example, a person would not describe their pet dog as disorganized or thorough). According to the theory of infrahumanization, an ingroup (e.g., the Spanish) may use HU emotions to describe themselves, but only use primary emotions to describe an outgroup (e.g., Catalonians). By humanizing themselves through the attribution of HU emotions, the Spanish create a disparity between the levels of humanness attributed to themselves (the ingroup) and Catalonians (the outgroup), thus implicitly dehumanizing the outgroup. Furthermore, infrahumanization can arise regardless of the valence of emotions. HU emotions are simply any emotions that are judged to exclusively arise in humans, and they may be either positive (e.g., broad-minded) or negative (e.g., stingy). As a result, infrahumanization can arise even in cases of a negative evaluation of the ingroup – that is, the ingroup could use negative, HU emotions to describe themselves while still using primary emotions to describe the outgroup (Haslam & Loughnan, 2013).

This concept of infrahumanization in turn informed Haslam’s (2006) dual model of dehumanization. In this account, Haslam proposes that there are in fact two types of humanness, and consequently two types of dehumanization: animalistic and mechanistic dehumanization. The former occurs when the aforementioned HU characteristics (e.g., amusement, intelligence,
and skepticism) are denied to individuals, reducing the distinction between them and animals (e.g., by implying that they possess a lack of morals or self-control). The second form of dehumanization occurs when human nature (HN) characteristics (e.g., openness to experience, emotionality, and agency), which are understood to be features that are typical of humans, are denied to others, which reduces the distinction between them and machines (e.g., by implying that they are emotionally cold). Haslam’s (2006) dual model of dehumanization expands on Leyens et al.’s (2001) model of infrahumanization in two ways. First, whereas Leyens et al. (2001) proposed that infrahumanization arises in regard to HU emotions, Haslam (2006) has expanded this to phenomenon to apply to the both emotions and personality characteristics (e.g., stinginess, warmth), and second, whereas Leyens et al. (2001) only considered HU emotions, Haslam (2006) considers both HU and HN traits.

While Haslam (2006) is the first to posit a dual model of dehumanization, other researchers have proposed parallel models that lend credence to the constructs identified by Haslam (2006). First, Gray, Gray, and Wegner (2007) identified that mind perception—that is, the recognition of another being’s mind—occurs along the dimensions of agency and experience. These dimensions map onto Haslam’s (2006) model of dehumanization, such that both the HU and agency dimensions account for a person’s ability to act with self-control and forethought, and the HN and experience dimensions account for a person’s consciousness and ability to experience emotions (Haque & Waytz, 2012). Second, Fiske, Cuddy, and Glick (2006) identify the dimensions of warmth and competence as the two universal dimensions of social cognition (i.e., the dimensions along which that we perceive and interact with others), which again map onto Haslam’s (2006) model. Both the HU and competence dimensions account for perceptions
of a person’s intelligence and skills, while the HN and competence dimensions account for perceptions of a person’s ability to act intentionally.

Although Haslam’s (2006) model is the only to explicitly use its measures to assess dehumanization, both Gray et al. (2007) and Fiske et al.’s (2006) models have been used as measures of dehumanization. Because the possession of a mind is an essential aspect of what makes one human, within the theory of mind perception dehumanization has been operationalized as reduced ratings along the dimensions of agency and experience (Cameron, Harris, & Payne, 2016). Using a neuroscientific approach within the theory of social cognition, Harris and Fiske (2006) found that individuals perceived as low on both the dimensions of competence and warmth were dehumanized (operationalized as reduced activity in the medial prefrontal cortex, an area linked to social cognition, of participants viewing images of these individuals).

Admittedly, these theories do not overlap perfectly. For example, while Gray et al.’s (2007) agency dimension accounts for all aspects of agentic action (e.g., self-control, thought, and planning), aspects of agentic action are divided between Haslam’s two dimensions: while the HU dimension accounts for self-control, the HN dimension accounts for agency more broadly. Nonetheless, these parallel models support Haslam’s (2006) model of dehumanization in two ways. First, while each of these theories proposes to measure a unique aspect of social perception, they all nonetheless measure social perception along two dimensions. Second, each of these theories identifies, in one way or another, that perceptions of a person’s ability to act agentially and experience emotions comprise an important part of social perception. Thus, the mind perception and social cognition theories lend support to both Haslam’s (2006) hypothesis
that dehumanization occurs along two dimensions, as well as to the characteristics that he proposes comprise each dimension.

Having addressed the theoretical underpinnings of Haslam’s (2006) model of dehumanization, one can now consider how it relates to Malle et al.’s (2014) path model of blame. To begin, these models overlap in the way that they take into account an individual’s intentionality and capacity. Haslam’s (2006) HN dimension maps onto Malle et al.’s (2014) intentionality component of the path model (such that intentionality is the principal component of agentic action; Bandura, 2001), while the HU dimension relates to the capacity component of the path model, in that both consider a person’s self-control. Based upon these parallels, the same factors that may drive mitigated blame attribution may in turn augment dehumanization. According to the path model, lower intentionality and capacity judgments should in turn mitigate blame attribution, but according to Haslam’s (2006) model these judgments should also mitigate perceived humanness, thus augmenting dehumanization.

To address these possibilities, Study 1 examined how people perceive the intentionality and capacity of an individual with ASD and an NT individual with regard to moral violations, as well as how these judgments affected the amount of blame attributed to these individuals and the extent to which they were dehumanized. Based on the way that intentionality and capacity judgments relate to theories of blame and dehumanization, the following predictions were developed.

1. Participants will perceive less intentionality and capacity in a target with ASD than an NT target when a moral violation is related to ASD. Consequently:
   a. The individual with ASD will be blamed less than the NT person, but
   b. The individual with ASD will be dehumanized more than the NT person.
Method

Participants

Participants (N = 181) were recruited from Macalester College’s student body and a forum called, “SampleSize,” on the website Reddit (www.reddit.com/r/samplesize). Participants were not compensated for their time.

Design

This experiment examined how targets with ASD and Type 1 Diabetes (T1D) were differentially assigned blame and dehumanized. This resulted in a 2 (target type: ASD or T1D) x 3 (norm violation type: social norm violation, problem behavior, and non-ASD related norm violation) mixed factorial design, in which the former variable was manipulated between-participants and the latter was manipulated within-participants. The social norm violation and problem behavior vignettes were written so that the norm violation in each vignette could be plausibly affected by ASD. The unrelated norm violation was included as a control condition, in order to assess whether participants uniformly mitigate blame attribution, regardless of whether the norm violation is linked to deficits related to ASD.

Materials

This experiment consisted of three principal sections. In the first, all participants read a short target description of the protagonist of the vignettes, Tim, which described his interests and daily routine. In addition to this description, participants in the ASD description read a short description of ASD, outlining the main aspects of the disorder (e.g., “First, individuals with ASD may be more dependent on routines, sensitive to change, and have more repetitive behaviors or interests than typically developing people… Second, individuals with ASD often have social deficits. These are often related to problems with their “theory of mind”, which is our ability to...
imagine or take someone else’s point of view”) as described in the DSM-V (American Psychiatric Association, 2013). In addition to the description of the protagonist, participants in the T1D condition also read a short description of T1D that outlined the main aspects of the disease (e.g., Tim also has Type 1 diabetes (T1D), which is an autoimmune disease that he was born with. T1D causes the body to destroy the cells that produce insulin, which is a hormone that enables people to get energy from food) as described by the World Health Organization (2010; Appendix A).

The second section contained three moral violation vignettes. The first depicted the target committing a social norm violation (i.e., talking about winning a competition to a student that lost the same competition), the second depicted him exhibiting problem behavior (i.e., becoming aggressive in response to a sudden schedule change), and the third depicted him planning and executing a malicious action (i.e., hiding a teammate’s soccer gear in order to get more playing time; Appendix B).

The final section contained a dehumanization questionnaire that asked participants to rate the target on 8 different personality traits (Appendix D). Four of these traits were HU characteristics (e.g., broad-minded, stingy) and 4 were HN characteristics (active, impatient; Haslam, Bain, Douge, Lee, and Bastian, 2005). The ratings for the personality traits in each of these categories were averaged to achieve a mean rating for each type of humanness. An equal number of positive (i.e., broad-minded, fun-loving, impatient, thorough) and negative traits (i.e., disorganized, shy, active, stingy) were presented.

Procedure

Recruited participants followed a hyperlink to Qualtrics, a survey hosting website on which the experiment took place. Informed consent was obtained for all participants. Participants
were randomly assigned to either the ASD or T1D condition, and read the corresponding description of the target. Following this, all participants read all three vignettes (the order of presentation was randomized for each participant) and responded to a unique set of three questions after reading each (Appendix C). The first two questions following each vignette were the same, and they asked the participants how much blame they believed the target deserved, and how intentionally they believe he acted. The third question assessed the same overarching construct (i.e., the capacity the target had to act otherwise; Malle et al., 2014), although it was tailored to the type of norm violation depicted in the corresponding vignette. Specifically, in the social norm violation condition, participants were asked whether the target knew that what he said would hurt the other students’ feelings. In the problem behavior condition and non-ASD related norm violation condition, participants were asked whether the target was in control of his behavior.

After participants completed this section, they completed the dehumanization questionnaire as well as several other demographic questions. Following this, participants were debriefed and thanked for their participation. Throughout the experiment, participants were unable to revisit text passages and questionnaires after they had already read or answered them.

**Results**

Two-hundred and fifty responses were initially recorded, but many participants failed to complete significant portions of the survey. Consequently, if participants failed to answer six or more questions (i.e., roughly a third of the survey), they were excluded from further analysis. This led to the exclusion of 69 participants, leaving 181 participants in the final sample.

The personality characteristics on the dehumanization questionnaire were classified as representing either the HN or HU dimension of humanness. A Cronbach’s alpha was computed
for the characteristics in each of these categories, but neither the characteristics composing the
HN ($\alpha = -.05$) or the HU ($\alpha = -.11$) constructs reached acceptable levels of internal consistency.
Following this, all eight personality characteristics were aggregated to be used as a composite
measure of the target’s “humanity,” but this grouping once again failed to achieve internal
consistency ($\alpha = .18$) Due to the inability to create an internally consistent measure of
participants’ perceptions of the target’s humanity, these ratings were excluded from further
statistical analysis.

To test the hypothesis that blame, intentionality, and capacity ratings would be lower for
a target with ASD than for an NT target, independent groups $t$-tests were conducted.
Participants’ ratings differed significantly in the expected direction for the two vignettes in
which the target’s behavior was plausibly influenced by aspects of ASD. In the social norm
violation vignette (Figure 1), participants in the ASD condition blamed the target less ($M = 2.90$
$SD = 1.39$) than did participants in the T1D condition ($M = 4.70$, $SD = 1.60$), $t(181) = -7.97$, $p <$
.001, $d = -1.20$. Similarly, perceived intentionality was lower in the ASD condition ($M = 1.49,$
$SD = .74$) than the T1D condition ($M = 3.25$, $SD = 1.49$, $t(181) = -9.73$, $p < .001$, $d = -1.50$) as
was perceived capacity (ASD: $M = 1.70$, $SD = .83$; T1D: $M = 3.46$, $SD = 1.47$, $t(181) = -9.96$, $p$
$< .001$, $d = -1.48$. This pattern of results was replicated in the problem behavior vignette (Figure
2). Once again, participants in the ASD condition blamed the target less ($M = 3.69$ $SD = 1.53$)
than participants in the T1D condition ($M = 6.08$, $SD = .99$), $t(181) = -12.75$, $p < .001$, $d = -2.21$,
and also perceived less intentionality (ASD: $M = 2.96$, $SD = 1.38$; T1D: $M = 5.39$, $SD = 1.56$),
$t(181) = -9.62$, $p < .001$, $d = -1.43$) and less capacity (ASD: $M = 2.96$, $SD = 1.38$; T1D: $M$
$= 5.11$, $SD = 1.77$, $t(181) = -8.96$, $p < .001$, $d = -1.36$). Contrary to the proposed hypothesis, this
pattern of results was also replicated in the control vignette (Figure 3), in which the target
sabotages another player on his soccer team. As in the other two vignettes, participants in the ASD condition again blamed the target less ($M = 5.58, SD = 1.50$) than participants in the T1D condition ($M = 5.58, SD = 1.50, t(181) = -7.60, p < .001, d = -1.08$), and perceived less intentionality (ASD: $M = 5.58, SD = 1.50$; T1D: $M = 6.76, SD = .71, t(181) = -7.06, p < .001, d = -1.01$) and capacity (ASD: $M = 5.06, SD = 1.61$; T1D: $M = 6.65, SD = .71, t(181) = -8.92, p < .001, d = -1.28$).

In sum, participants perceived that the target with ASD acted with less intentionality and capacity, and deserved less blame, than the NT target across all scenarios. To further explore the relationship between these variables, and test whether the relationship between condition and blame was mediated by perceived intentionality or capacity, bivariate correlations and mediation analyses were calculated.

In the social norm violation vignette, both intentionality ($r = .59$) and capacity ($r = .58$) were correlated with blame, both $ps < .001$. The effect of condition was mediated by both perceived intentionality [95% CI of indirect effect: .33, 1.16] and capacity [95% CI of indirect effect: .31, 1.13]. After accounting for the joint effects of the mediators, the original direct effect of condition on blame ($t = 7.97, p < .001$) was reduced to marginal significance ($t = 1.68, p = .09$). Parallel correlations and mediation analyses were also calculated for participants’ ratings of the target in the problem behavior and control vignettes. In the PB vignette, blame was correlated with perceived intentionality ($r = .58$) and capacity ($r = .55$), both $ps < .001$. The relationship between condition and blame was mediated by the perceived intentionality [95% CI of indirect effect: .66, 1.38] and capacity [95% CI of indirect effect: .01, .52]. After accounting for the joint effects of these mediators, the original direct effect of condition on blame ($t = 12.75, p < .001$) remained significant but was mitigated ($t = 6.42, p < .001$). In the control vignette, blame was
again correlated with perceived intentionality \((r = .47)\) and capacity \((r = .55)\), both \(p < .001\).

Once again, the relationship between condition and blame was mediated by perceived intentionality \([95\% \text{ CI of indirect effect: .28, .86}]\) and capacity \([95\% \text{ CI of indirect effect: .16, .69}]\). After accounting for these mediators, the original direct effect of condition on blame attribution \((t = 7.60, p < .001)\) was reduced to nonsignificance \((t = 1.33, p = .19)\), indicating that this association was fully mediated by perceived intentionality and capacity.

**Discussion**

In this study, participants attributed less blame to a target with ASD, and perceived that he acted with less intentionality and capacity, than an NT target (i.e., the target with T1D) in all vignettes. Furthermore, this association between target type and blame was partially mediated by perceived intentionality and capacity in the social norm violation and problem behavior vignettes, and fully mediated by these factors in the control vignette. These results support the first hypothesis: that a target with ASD will receive less blame than an NT target for committing an identical norm violation when the norm violation is putatively related to ASD. Unexpectedly, this pattern also emerged in the control vignette, in which the target’s norm violation was putatively unrelated to aspects of ASD.

Unfortunately, the HN and HU characteristics adapted from Haslam et al.’s (2006) study failed to reach internal consistency, either as independent constructs or when grouped together, as a measure of total humanity. As a result, it was not possible to run analyses testing the second hypothesis: that participants would dehumanize the target with ASD target more than the NT target. The poor observed internal consistency might stem from the design of the dehumanization questionnaire, as well as the source from which the characteristics were taken. In Haslam et al.’s (2005) study, 32 characteristics were categorized as high or low on dimensions of desirability,
human nature, or human uniqueness. To obtain these groupings, the researchers asked
participants to rate each characteristic on each of these three dimensions, and then conducted t-
tests to determine which characteristics were rated higher on a given dimension than on the other
two dimensions. Importantly, the authors did not calculate measures of internal consistency for
the set of characteristics that composed each dimension. This limitation, paired with the fact that
the current study selected only four traits to comprise each construct (trimmed down from the 16
characteristic per group provided in Haslam et al.’s (2006) study), likely contributed to our
failure to obtain internal consistency. Future research may benefit both from the selection of a
larger number of characteristics to comprise each construct, as well the use of pre-testing to
ensure that the characteristics comprising each construct reaches a reliable level of internal
consistency before employing them in a complete study.

Given that blame and punishment are positively associated (Carlsmith et al., 2002) and
share similar theoretical frameworks (Darley & Pittman, 2003; Malle et al., 2014), and that
people with ASD are assigned mitigated, or alternative, punishments (Carr et al., 2002; Safran &
Oswald, 2003), we hypothesized that blame may likewise be mitigated for individuals with ASD.
The results of Study 1 show that people with ASD are indeed blamed less than their NT
counterparts, and that this effect is mediated by perceived intentionality and capacity, consistent
with Malle et al.’s (2014) path model.

These results are of further interest in relation to Gray, Young, and Waytz’s (2012)
theory of dyadic morality (TDM). This theory proposes that morality is perceived along two
dimensions: moral agency (i.e., a person’s capacity to execute moral actions) and moral patiency
(i.e., a person’s capacity to experience moral actions). While every moral situation is comprised
of a moral agent and a moral patient who are different from each other, each person’s general
moral status is perceived along both of these dimensions simultaneously. When a person’s moral status is considered, Gray et al.’ (2012) propose that these dimensions are inversely related; the higher a person is perceived along one dimension (e.g., moral agency) the lower they are perceived along the other (e.g., moral patiency), such that people are “morally typecast” into one role or another (Gray & Wegner, 2009). For example, Gray and Wegner (2009) found that superheroes are typecast as moral agents, and are consequently rated as low in moral patiency, whereas civilians are not typecast as either role, and are thus perceived along these dimensions in equal measure. In this way, while both a superhero and a civilian occupy the role of a moral patient if they are tortured, participants perceive that a superhero feels less pain than a civilian; that is, participants perceive the superhero as less of a moral patient than the civilian due to the superhero being typecast as a moral agent.

In Study 1, a target with ASD, as compared to an NT target, was perceived to act with less intentionality and capacity, and consequently to deserve less blame. Blame and praise are used as measures of a target’s moral agency; thus, these results suggest that the target with ASD was perceived as a lesser moral agent. According to Gray et al.’s (2012) TDM, these reduced perceptions of moral agency should correspondingly augment perceptions of moral patiency, due to the inverse relationship between these dimensions. That is, the target with ASD should be typecast as a moral patient. Extant literature analyzing Anglophone media representation of people with ASD offers preliminary support for this hypothesis. In their review of British newspapers from 1999-2008, Huws and Jones (2010) found that people with ASD were most often described as “victims” of their disorder, or as “suffering.” These descriptions portray people with ASD as possessing low levels of moral agency (victimization assigns agency to the disorder rather than the individual) and augmented levels of moral patiency (moral patiency is
operationalized as suffering). This pattern of representation also emerged in a review of Australian newspaper articles that addressed portrayals of autism spectrum disorders (Jones & Harwood, 2008). The authors found that people with ASD were often described as maltreated and neglected, thus again typecasting them as moral patients. Interestingly, these authors also observed conflicting forms of representation, such that people with ASD were also portrayed as violent, uncontrollable, and a burden to their friends and family, traits which are linked to moral agency. Even in this case, individuals with ASD may still be portrayed as moral patients, as ASD is presented as the cause of their actions (e.g., aggression), and the individuals themselves as subject to the control of this disorder.

Gray et al.’s (2012) TDM predicts that the mitigated perceived moral agency of people with ASD in Study 1 should typecast them as moral patients, and thus augment perceptions of their moral patiency. Further supporting this hypothesis, reviews of Anglophone print media provide initial evidence that people with ASD may indeed be typecast as moral patients (Huws & Jones, 2010; Jones & Harwood, 2008). Study 2 sought to experimentally ascertain whether people with ASD are indeed typecast as moral patients.

**Study 2**

Gray et al.’s (2012) TDM propose that “mind perception is the essence of morality” (p. 103), and consequently base their dimensions of moral agency and patiency respectively off of the dimensions of agency and experience proposed in Gray et al.’s (2007) theory of mind perception (TMP). According to the TMP (Gray et al., 2007), we perceive others’ minds along the dimension of agency (i.e., their capacity to plan, hold goals, and act intentionally) and experience (i.e., their ability to experience emotions, pleasure, and suffering), which are similar to Gray et al.’s (2012) dimensions of moral agency and patiency. Beyond the similarity in their
theoretical claims, these two theories are also similar with respect to the measures used to assess them. On Gray et al.’s (2007) Mind Scale questionnaire (which includes agency and experience subscales) participants rate a target’s ability to experience pain and pleasure, both of which are also used as measures of a target’s moral patiency. While agency and moral agency are not assessed with identical items, they nonetheless have substantial overlap. The TDM operationalizes moral agency as a target’s ability to incur blame and praise, while the TMP operationalizes agency as a target’s ability to plan and exert self-control—two traits that partially mediated blame attribution in Study 1. While these theories purport to measure different aspects of an individual (i.e., either their mind or their moral status), the theoretical and operational similarities between them invites an investigation of whether they differ substantively from one another.

The most notable difference between these theories is their conceptualization of the relationship between their respective dimensions. The TDM posits that moral agency and patiency are inversely related to one another, such that the greater an individual is perceived along one dimension, the lesser they are perceived along the other. In contrast to this, the TMP posits that these dimensions are independent of one another; that is, an individual’s rating along one dimension does not affect their rating along the other. Thus, if an individual receives elevated agency or moral agency ratings, according to the TDM this will lead to mitigated experience or moral patiency ratings, while according to the TMP this will have no effect on experience or moral patiency ratings. This apparent conflict has even been noted by Gray and Wegner (2009), whose experimental work on moral typecasting demonstrated the inverse relationship between moral agency and patiency posited by the TDM. These authors propose that this conflict is not indicative of a fundamental disagreement between these theories, but rather
that experience and agency can relate to each other along two broader dimensions: a mind perception dimension (where agency and experience are independent) and a moral perception dimension (where moral agency and patience are inversely related).

Indeed, experimental evidence supports both the conceptualization of agency and experience constituting a mind and experience versus agency constituting an individual’s moral status. In support of their TMP, Gray et al. (2007) identified agents whose ratings along the dimensions of agency and experience demonstrate every possible relationship between these two dimensions. For example, an average human man and woman are rated high on both agency and experience; God is rated high on agency but low on experience; a baby is rated low on agency and high on experience; and a dead man is rated low on both agency and experience. In support of their TDM, Gray and Wegner (2009) conducted several studies that consistently found that an agent’s moral agency and patience ratings are inversely related. For example, a child is typecast as a moral patient and consequently receives elevated moral patience ratings and mitigated moral agency ratings; conversely, a superhero is typecast as a moral agent and subsequently receives elevated moral agency ratings and mitigated moral patience ratings. Furthermore, these authors found that if a neutral target (e.g., an average man) acts as a moral agent or patient in one vignette, he will be typecast as such in future vignettes. Thus, one must ask how the dimensions of agency and experience, and moral agency and patience, can be theoretically and operationally similar, yet receive experimental support for the competing claims that they are independent of one another and inversely related to each other?

Importantly, the TDM and TMP appear to differ regarding the context in which they consider an agent. The experimental evidence demonstrating the inverse relationship between the dimensions of moral agency and patience (Gray & Wegner, 2009), proposed by the TDM, has
been collected with regard to specific moral situations. For example, Gray and Wegner (2009) asked participants to read about a person killing a man, and then rate him along the dimensions of moral agency and patiency. In this case, the man was perceived as high in moral agency and low in moral patiency, which the authors take as evidence that he was morally typecast as a moral agent based upon his behavior in the vignette. In contrast to this, the experimental evidence demonstrating that the dimensions of agency and experience are independent of one another, as proposed by the TMP (Gray et al., 2007), has been collected with regard to an agent’s general traits. In Gray et al.’s (2007) study, participants read brief descriptions of a target (e.g., that the target had been in a persistent vegetative state (PVS) for the past six months or that the target was a middle-aged woman who works at an advertising agency) and then rated them along the items on the Mind Scale; participants did not consider any specific behaviors that the agent performed. In this way, it appears that while the TDM and TMP use identical dimensions, albeit with nominal differences (i.e., agency or moral agency, and experience or moral patiency), these dimensions relate differently depending on the context in which they are considered. Specifically, we hypothesize that when a target is considered in a general context, the dimensions accounting for agency and experience are independent of one another (in line with the TMP), while if an agent is considered in a specific context, these dimensions are inversely related to one another (in line with the TDM). This prediction will hereafter be referred to as the context hypothesis.

This context hypothesis reconciles the competing predictions generated by the TMP and TDM concerning how an individual with ASD will be perceived along the dimensions of agency and experience, and moral agency and patiency. To understand how this functions, one can first consider the predictions generated by each theory, and then consider how these predictions may
be moderated by context. As mentioned, the dimensions of agency and experience proposed by the TMP have been used as a measure of dehumanization (Cameron et al., 2016), such that lower ratings on each of these dimensions indicates reduced overall mind perception and thus dehumanization. ASD is a stigmatized disorder (Obeid et al., 2015) and stigma has been identified as a factor that leads to dehumanization (Cameron et al., 2016). Thus, in line with the TMP, one predicts that the stigma related to ASD will lead to a target with ASD being dehumanized, operationalized as the receipt of lower ratings along the agency and experience dimensions. The TDM posits that the dimensions of moral agency and patiency are inversely related. Thus, based on the finding that a target with ASD was perceived as less of a moral agent than an NT target in Study 1 (operationalized as lower ratings of blame and intentionality), the TMD predicts that the target with ASD will consequently be perceived as a greater moral patient (operationalized as his increased capacity to experience suffering or pleasure).

According to the context hypothesis, the competing predictions proposed by the TDM and TMP can be reconciled by considering the different contexts in which the target is considered by participants. Specifically, the way that the dimensions of agency and experience, and moral agency and patiency, relate to each other can be predicted by the TMP in the general context condition, and by the TDM in the specific context condition. Thus, one predicts that if a target with ASD is considered in a general context he will receive mitigated agency and moral agency, and experience and patiency, ratings (hereafter referred to for this study as the general prediction). One also predicts that if he is considered in a specific context he will receive mitigated agency and moral agency ratings, but elevated experience and moral patiency ratings, (hereafter referred to for this study as the specific prediction).

**Method**
Participants

Participants ($N = 640$) were recruited for $$.30 on MTurk, a marketplace run by Amazon on which individuals can complete surveys or other tasks for monetary compensation.

Design

This experiment examined how participants assigned blame, and perceived the agency and experience, and moral agency and patiency, of a target with autism spectrum disorder (ASD) or type 1 diabetes (T1D), based upon their knowledge of either only a general description of the target, or also their additional knowledge that the target executed a moral action. This resulted in a 2 (target type: ASD or T1D) x 2 (context: general or specific) x 2 (perception type: mind or moral) between subjects design.

Materials

The target descriptions (ASD and T1D) were identical to those used in Study 1 (Appendix A). The moral violation vignettes (social norm and control) likewise were identical to those used in Study 1 (Appendix B). The Mind Scale questionnaire (Gray et al., 2007; Appendix E) consisted of items assessing the extent to which the target possessed a number of mental traits along the dimensions of agency (self-control, morality, emotion recognition, memory, planning, communication, thought) and experience (hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy). Finally, five moral perception items assessed the target’s moral agency (capability to deserve blame or praise for his actions) and moral patiency (capability to experience suffering or pleasure; Appendix F).

Procedure

Recruited participants followed a hyperlink to Qualtrics, a survey hosting website on which the experiment took place. Informed consent was obtained for all participants. Participants
were randomly assigned to either the ASD or T1D condition, and then read the corresponding
description of the target.

Participants were then randomly assigned to either the general or specific context
condition, as well as to the mind perception or moral perception condition. After reading the
target description, participants in the specific context condition were randomly assigned to read
one of the two moral agency vignettes. Then they provided ratings about the target, completing
either the mind perception items (Mind Scale survey: Gray et al., 2007) or the moral perception
items, depending on their randomly assigned condition. Participants in the general context
condition did not read about a specific behavior; they provided target ratings (mind perception or
moral perception) immediately after reading the target description.

Upon their completion of their respective questionnaires participants were debriefed and
dismissed. Throughout the experiment, participants were unable to revisit text passages and
questionnaires after they had already read or answered them.

Results

Reliability analyses were run for the items used in the Mind Scale (Gray et al., 2007) and
the Moral Perception questionnaire. The items comprising the agency dimension of the Mind
Scale had excellent internal consistency ($\alpha = .84$), as did the items comprising the experience
dimension ($\alpha = .87$); consequently, the items comprising these dimensions were averaged to form
agency and experience scores. The items comprising the agency subscale of the Moral
Perception questionnaire (moral responsibility, blame, and praise) demonstrated poor internal
consistency ($\alpha = .31$), which was due to poor cohesion between praise ratings and the other two
variables. Blame and moral responsibility ratings were strongly correlated ($r = .68, p < .001$), as
were the items assessing perceived moral patiency (capacity for suffering and pleasure; $r = .75, p$
<.001). Thus, blame and moral responsibility ratings were averaged to form a moral agency score, and capacity for suffering and pleasure ratings were averaged to form a moral patiency score.

First, we tested the general context hypothesis—namely, that when reading only a general description of the target, participants would view the target with ASD as lower in agency and experience (and, similarly, as lower in moral agency and patiency) than the NT target. In support of this hypothesis, in the general context condition the target with ASD received lower agency ratings ($M = 3.00, SD = .63$) than the NT target ($M = 3.81, SD = .72$), $t(159) = -7.59, p < .001, d = -1.20$. Experience ratings, however, did not differ by target (ASD: $M = 3.37, SD = .75$; NT: $M = 3.45, SD = .68$) $t(159) = -.68, p = .50, d = -.11$ Revealing a parallel pattern, the target with ASD was seen as less of a moral agent ($M = 3.52, SD = .95$) than the NT target ($M = 3.98, SD = .97$, $t(161) = -3.04, p = .003, d = -.48$), but moral patiency ratings did not differ by target type (ASD: $M = 4.12, SD = .92$; NT: $M = 4.09, SD = 1.27$), $t(161) = .15, p = .88, d = .03$ (Figure 4).

Then, we tested our specific context prediction—namely that when reading about a target’s specific moral violation, participants would view the target with ASD as lower in moral agency but higher in patiency (and, similarly, as lower in agency but higher in experience) than an NT target. Restricting the analysis to the specific context condition, the target with ASD was indeed seen as lower in moral agency ($M = 4.02, SD = 1.45$) than the NT target ($M = 4.56, SD = 1.14$, $t(161) = -2.66, p = .01, d = -.41$), but moral patiency ratings did not differ by target type (ASD: $M = 3.90, SD = 1.55$; NT: $M = 3.86, SD = 1.23$), $t(161) = .18, p = .86, d = .03$. Neither ratings of agency (ASD: $M = 2.83, SD = .71$; NT: $M = 2.79, SD = .67$) nor ratings of experience (ASD: $M = 3.10, SD = .70$; NT: $M = 2.95, SD = .72$) differed by target type, $t$s$(159) < 1.30, ps > .10, ds < .3$ (Figure 5).
To test these ratings across context condition, factorial ANOVAs were run predicting agency, experience, moral agency, and patiency scores by target type, context type, and a target x context type interaction. A target x context type interaction was not observed for experience, moral agency, or patiency scores (all $F$s(327) < 2, $ps > .15$), but this interaction was observed for agency scores, $F(327) = 31.44, p < .001$). While the NT target received significantly higher agency scores in the general context ($M = 3.81, SD = .73$) than the target with ASD ($M = 3.00, SD = .63, t(159) = 7.59, p < .001, d = 1.19$), their agency ratings did not differ in the specific context (NT: $M = 2.79, SD = .71$; ASD: $M = 2.83, SD = .67, t(159) = -.41, p = .68, d = -.06$).

To test whether context moderated the relationship between agency and experience (or between moral agency and patiency), as proposed by the context hypothesis, generalized linear models were created. For the Mind Scale, agency ratings were regressed on experience ratings and the experience x context interaction. Experience ratings strongly predicted agency ratings ($\beta = .73, p < .001$), but the experience x context interaction was not a significant predictor ($\beta = .01, p = .90$). Thus, context did not moderate the relationship between agency and experience ratings. For the Moral Perception questionnaire, moral agency ratings were regressed on moral patiency ratings and the moral patiency x context interaction. Here, moral patiency ratings did not predict moral agency ratings ($\beta = .18, p = .83$), but there was a significant moral patiency x context interaction ($\beta = .25, p = .025$), indicating that these scores related differently to each other depending on context condition. In the general context, moral agency and patiency ratings were more strongly correlated ($r = .48, p < .001$) than in the specific context ($r = .20, p < .001$).

**Discussion**

According to the context hypothesis, we predicted that the relationship between perceived agency and experience would mirror that between perceived moral agency and moral patiency,
across each context condition (either general or specific) in this study. Our general context prediction stated that the target with ASD would be assigned lower agency and moral agency, and experience and patiency, ratings in the general context (thus indicating dehumanization). Our specific context prediction stated that the target would receive lower agency and moral agency, but elevated experience and patiency, ratings in the specific context. Each of these predictions received partial support.

In the general context condition, the target with ASD was assigned lower agency and moral agency ratings than the NT target but the experience and patiency ratings for the two targets did not differ significantly, thus only partially confirming our general context prediction. Nonetheless, agency and moral agency, and experience and patiency, ratings mirrored each other, as predicted by the context hypothesis. In the specific context condition, the target with ASD received lower moral agency ratings than the NT target, supporting our specific context prediction, although agency, experience, and patiency ratings did not differ by target as predicted. While experience and patiency ratings mirrored each other in this context as predicted by the context hypothesis, agency and moral agency ratings did not. Finally, a patiency x context interaction predicted moral agency scores, such that moral agency and patiency scores were more strongly correlated in the general context than the specific context. Although the prediction that these scores would be positively correlated in the general context and inversely correlated in the specific context was not confirmed, the observed correlations nonetheless differed in the expected direction (such that agency and experience scores demonstrated a stronger positive correlation than moral agency and patiency scores). In sum, these results generate only partial support for the context hypothesis. While the relationship between moral agency and patiency ratings mirrored that between agency and experience ratings in the general context, this pattern
was not observed in the specific context. Additionally, while context moderated the relationship between moral agency and patiency scores, it did not moderate the relationship between agency and experience scores.

When our hypotheses were unconfirmed it was due to a lack of a significant difference between scores by target type, rather than differences arising opposite of the predicted direction. Consequently, while the failure to observe significant differences may indeed be due to fault with the hypotheses themselves, it may also be due to a lack of statistical power. Specifically, the difference between the two targets may not have been significant enough to reveal differences in their ratings along each dimension. The experimental evidence supporting the TMP (Gray et al., 2007) and TDM (Gray & Wegner, 2009) used a diverse range of targets intended to test the extremes of each theory. For example, in Gray and Wegner’s (2009) studies on moral typecasting they asked participants to consider two targets with extreme, salient differences. In one study, participants rated individuals famous for either good or bad moral actions (e.g., Mother Theresa or Ted Bundy, both typecast as moral agents) and normal citizens; in another they rated young children (typecast as moral patients) and adults.

While developing their theory of mind perception, Gray et al. (2007) collected information on a variety of agents, not even all of whom were human (e.g., a dead man, a dog, and God), who represented the extremes of the agency and experience ratings that one can receive (e.g., a dead man is rated low on agency and experience, a dog is rated low on agency and high on experience, God is rated high on agency and low on experience). Within normally functioning human targets, there was significantly less diversity in agency and experience ratings. In fact, all human targets with normal cognitive functioning (i.e., a child and two adults) received near maximum experience ratings, with only their agency ratings differing. In the
present study, participants rated one of two middle school students who had the same interests and hobbies. The only difference between them was that one had ASD while the other had T1D. Furthermore, the description of ASD saliently implicated the target’s ability to act agentically (e.g., “[people with ASD] sometimes display aggressive or destructive behaviors when they become angry” and “as a result of these deficits, individuals with ASD may make social faux pas or say something offensive to others”; Appendix A) while it addressed traits related to experience/patiency to a much lesser extent. Relatedly, the most consistent differences found between the targets pertained to their agency and moral agency scores: the target with ASD received lower agency and moral agency scores in the general context, and lower moral agency scores in the specific context. The targets’ experience and moral patiency scores did not differ in either context. In this way, these two targets may have been too similar to expose differences regarding their perceived experience and patiency (as predicted by the TMP and TDM) even if these differences indeed exist.

Another potential limitation of this study pertains to the prediction that the target with ASD would receive lower agency and experience ratings (and moral agency and patiency ratings) if considered in a general context. This prediction was based upon the fact that ASD is a stigmatized disorder (Obeid et al., 2015) and stigma is linked to dehumanization (Cameron et al., 2016) – operationalized as reduced agency and experience ratings on the Mind Scale. The stigma associated with ASD may not have been strong enough to lead to dehumanization, though. In Cameron et al.’s (2016) study, they only observed this relationship between stigma and dehumanization when asking participants to consider helping a homeless drug addict (a highly stigmatized social category). Notably, when the researchers asked participants to simply read about (rather than consider helping) this target, or to consider helping a less stigmatized target
(i.e., an individual who is homeless due to the cost of treatment for a medical condition), the link between stigma and dehumanization was not present. In contrast to this, participants in the present study were simply asked to read about a target that has ASD, the stigma associated with which is much less than that associated with homelessness or drug addiction. Thus, perhaps the stigma associated with ASD was not strong enough or salient enough to lead to the dehumanization of a person with ASD.

This explanation is further supported by a consideration of stigma and dehumanization with regard to Fiske et al.’s (2006) model of social cognition. This model proposes that we perceive others along the dimensions of competence and warmth, which line up respectively with the dimensions of agency and experience (and moral agency and patiency). Harris and Fiske (2006) found that only people who are identified as low on both competence and warmth are dehumanized. Furthermore, Harris and Fiske (2006) note that while drug addicts are classified as low on both of these dimensions—thus corroborating Cameron et al.’s (2016) finding that a homeless drug addict is dehumanized—“disabled people” are rated low on competence but high on warmth. If one translates these ratings to the respective dimensions used by the TMP and TDM, a “disabled person” (e.g., a person with ASD) would be rated low on agency and moral agency, but high on experience and patiency. Thus, while ASD is indeed a stigmatized disorder, it may only affect perceptions of a person’s agency, leaving perceptions of their experience unaffected, explaining our failure to observe differences in the experience and patiency ratings assigned to targets in the general context condition of this study.

Study 3

The failure to obtain full support for our hypotheses (i.e., the context hypothesis, general context prediction, and specific context prediction) in Study 2 appeared to be due in part to the
insufficient strength of our experimental manipulation. To maximize the strength of our
manipulations, and to improve our fidelity to the methodology of the studies that first generated
evidence for the TMP and TDM, we drew more directly off of the design of these original
studies. Specifically, we used an adult woman and a young girl as our targets (originally used by
Gray et al., 2007) and we had participants read two vignettes either taken from, or based off of,
Gray and Wegner’s (2009) moral typecasting studies. These changes generated several new
predictions. Based upon the agency and experience ratings assigned to a child and woman in
Gray et al.’s (2007) study, one predicts that the child will receive lower agency ratings than the
woman in the general context condition, but that their experience scores will not differ (hereafter
referred to for this study as the general context prediction). The TDM generates two predictions
concerning typecasting in the specific context condition (hereafter referred to for this study as the
specific context prediction). First, in line with Grey and Wegner’s (2009) findings, the child
should be typecast as a moral patient, thus receiving lower moral agency ratings, but higher
moral patiency ratings, than the woman. Second, if participants first read about a target acting as
a moral agent, they should typecast her as such, and likewise if they first read about a target as
moral patient. Finally, the context hypothesis predicts that the relationship between agency and
experience ratings will mirror the relationship between moral agency and patiency ratings within
each context condition.

Method

Participants

Participants (N = 720) were recruited for $0.40 on MTurk.

Design
This experiment examined how participants assigned blame, and perceived the agency and experience, and moral agency and patiency, of a young girl and an adult woman, based upon their knowledge of either only a general description of the target, or also their additional knowledge that the target executed and experienced moral actions. This resulted in a 2 (target type: child or adult) x 2 (perception type: mind or moral) x 2 (order: moral agency first or moral patiency first) between subjects design. A single general context condition was also included, in which participants completed all of the above conditions as a within-subjects design.

Materials

The target descriptions were taken from Gray et al.’s (2007) study on mind perception (“Sharon Harvey, 38, works at an advertising agency in Chicago” and “Samantha is a five-year-old girl who lives with her parent and older sister Jennifer”). The first moral violation vignette was taken from Gray and Wegner’s (2009) paper proposing the moral typecasting hypothesis, while the second vignette was developed specifically for this experiment (Appendix G). The Mind Scale questionnaire (Gray et al., 2007; Appendix E), first used in Study 2, was again used. Similarly, the moral perception questions (Appendix H) were also adapted from Gray and Wegner’s (2009) study.

Procedure

Recruited participants followed a hyperlink to Qualtrics, a survey hosting website on which the experiment took place. Informed consent was obtained for all participants. Participants were randomly assigned to either the child or adult condition, and then read the corresponding description of the target.

Then, participants were randomly assigned to either the general or specific context condition, as well as to the mind perception or moral perception condition. After reading the
target description, participants in the specific context condition were then assigned to one of two orders in which they read the two moral vignettes. Participants in the moral agency first order condition read about the target acting as a moral agent in the first vignette and then as a moral patient in the second; participants in the moral patiency first order condition read about the target acting as a moral patient in the first vignette and as a moral agent in the second (Appendix G). Participants in the moral perception condition answered the moral perception questions after each vignette, while participants in the mind perception condition completed the mind perception questions (Mind Scale: Gray et al., 2007) after reading both vignettes. Participants in the general context condition did not read about specific behaviors; they provided target ratings (mind perception or moral perception) immediately after reading each target description.

Upon their completion of their respective questionnaires, participants were debriefed and dismissed. Throughout the experiment, participants were unable to revisit text passages and questionnaires after they had already read or answered them.

**Results**

The items comprising the agency dimension of the Mind Scale had excellent internal consistency ($\alpha = .91$) as did the items comprising the experience dimensions ($\alpha = .87$). The two moral agency items (perceived moral responsibility and intentionality) were moderately correlated ($r = .39$) and were consequently averaged to create a moral agency score. The perceived pain felt by the target was used as a direct measure of moral patiency.

According to the TMP prediction, in the general context condition the adult woman should be perceived as higher in agency than the child, but their experience scores should not differ. Indeed, the woman received high agency scores ($M = 3.84, SD = .77$) than the child ($M = 2.84, SD = .76, t(134) = -2.73, p < .001, d = 1.31$), while their experience scores did not differ
(woman: $M = 3.49$, $SD = .88$; child: $M = 3.37$, $SD = .62$), $t(134) = .90$, $p = .37$, $d = -.14$. The context hypothesis predicted that in the general context condition moral agency and patiency ratings would follow this same pattern. Confirming this prediction, the woman received higher moral agency ratings ($M = 5.87$, $SD = 1.07$) than the child ($M = 3.63$, $SD = 1.10$, $t(134) = 11.98$, $p < .001$, $d = 1.06$) while their moral patiency scores did not differ (woman: $M = 5.72$, $SD = 1.28$; child: $M = 5.54$, $SD = 1.64$), $t(134) = .70$, $p = .49$, $d = .12$ (Figure 6).

Although we did not hypothesize about a possible order effect, one emerged. As compared to participants who first rated the woman, those who first rated the child gave higher ratings on all four measures (agency, experience, moral agency, and moral patiency) for both targets all $ts(134) > 2.40$, all $ps < .05$. To test whether target type moderated any of these order effects, factorial ANOVAs were run predicting agency, experience, moral agency, and patiency ratings by order type, target type, and an order x target type interaction. No order x target type interaction was revealed, all $Fs(135) < 3$, all $ps > .10$, indicating that target type did not moderate these effects.

According to the specific context prediction, the child should be typecast as a moral patient in the specific context, resulting in lower moral agency ratings but higher moral patiency ratings than the woman target. In contrast to this hypothesis, neither moral agency ratings (child: $M = 4.61$, $SD = 1.53$; woman: $M = 4.57$, $SD = 1.56$), $t(334) = .07$, $p = .91$, $d = .03$), nor moral patiency ratings (child: $M = 4.40$, $SD = 1.05$; woman: $M = 4.33$, $SD = 1.08$), $t(334) = .46$, $p = .57$, $d = .07$) differed across target type. The specific context prediction further proposes that if targets are perceived to act as a moral agent or patient in one situation, they should be typecast as such in future situations. Accordingly, participants should rate the target as lower in moral patiency when they have previously read about the target acting as a moral agent. Likewise,
participants should rate the target as lower in moral agency when they have previously read
about the target as a moral patient. The results failed to confirm these hypotheses. Perceived
moral agency did not differ between the moral agency first ($M = 4.59, SD = 1.56$) and moral
patiency first order conditions ($M = 4.61, SD = 1.53, t(334) = -0.12, p = .91, d = -.01$). Perceived
patiency ratings also failed to differ between the agency first ($M = 4.33, SD = 1.08$) and patiency
first order ($M = 4.40, SD = 1.05, t(334) = -0.55, p = .56, d = -.07$).

According to the context hypothesis, this observed pattern of results should have also
been observed for agency and experience ratings provided by participants in the specific context
condition. Supporting our specific context prediction (that the woman would receive higher
agency ratings than the child), although failing to mirror the moral agency ratings (and failing to
support the context hypothesis), the woman was seen as having more agency ($M = 3.01, SD =
.82$) than the child ($M = 2.72, SD = .71, t(334) = 3.45, p < .001, d = .39$). Once again perceptions
of experience did not differ by target type (woman: $M = 3.12, SD = .72$; child: $M = 3.03, SD =
.67$), $t(334) = 1.15, p = .25, d = .13$. As with agency and experience ratings, neither agency nor
experience ratings showed order effects, both $t$s$(334) < 1.0$ (Figure 7).

To test whether context moderated the relationship between agency and experience (or
between moral agency and patiency), generalized linear models were created. For the mind scale,
agency ratings were regressed on experience ratings and an experience x context interaction.
Both experience ratings ($\beta = .85, p < .001$) predicted agency ratings and the experience x context
interaction ($\beta = -.05, p < .001$) predicted agency ratings. This interaction functioned such that
agency and experience ratings exhibited a strong correlation in the specific context ($r = .76, p <
.001$) than in the general context ($r = .62, p < .001$) For the moral perception items, moral agency
ratings were regressed on patiency ratings and a patiency x context interaction. While moral
patiency ratings ($\beta = .0.30$, $p < .001$) predicted moral agency ratings, the moral patiency x context interaction ($\beta = .008$, $p = .81$) did not, indicating that context did not moderate the relationship between these scores.

**Discussion**

In this study, the context hypothesis predicted that the relationship between agency and experience ratings would mirror that between moral agency and patiency ratings within each context condition. The general context hypothesis (based upon the TMP) predicted that the child would receive lower agency and moral agency ratings than the woman, but that their experience and patiency ratings would not differ. Indeed, this hypothesis was confirmed. Additionally, while an order effect was not predicted, one was observed, such that if participants rated the child first, they provided higher agency, experience, moral agency, and patiency ratings for *both* targets than if they rated the woman first. The specific context prediction (based upon the TDM) contained two predictions 1) that the child would be typecast as a moral patient, and thus receive lower moral agency and agency, and higher patiency and experience, ratings than the woman and 2) that if participants read about a target acting as a moral agent or patient in one vignette, they would typecast the target as such for subsequent vignettes. Each specific context prediction only received limited support. In the specific context condition, the child received lower agency scores than the woman, but their moral agency, patiency, and experience scores did not differ. Furthermore, only an order effect for agency scores was revealed (such that if participants read about the target acting agentically first they provided higher agency ratings and lower experience ratings, indicating that the target had been typecast as a moral agent). No such order effect was observed for moral agency, patiency, or experience scores.
In sum, while the predictions based upon the TMP were confirmed (thus corroborating the past experimental work of Gray et al., 2007), the predictions generated by the TDM received little support. In fact, one of the two pieces of supporting evidence for the TDM (that the girl received lower agency ratings than the woman in the specific context) cannot necessarily be attributed to processes unique to the TDM. Participants in both context conditions read the same target descriptions and answered the same agency items, while participants in the specific context condition additionally read vignettes depicting the targets in moral situations. Consequently, the fact that the girl received lower agency ratings in the specific and general context may stem from differences between the target descriptions of the two characters (taken from Gray et al.’s (2007) paper on the TMP) rather than the effect of reading moral vignettes (taken from Gray and Wegner’s (2009) paper supporting the moral typecasting hypothesis and TDM).

This failure to replicate Gray and Wegner’s (2009) findings is particularly surprising given the fidelity with which this study replicated the methodology of their original experimental work. This study directly used one vignette from their original set of studies, and the second vignette was closely modeled after the other vignettes used in these studies. Furthermore while target descriptions were taken from Gray et al.’s (2007) study, they closely mirrored those of Gray and Wegner (2009). For example, before reading vignettes, participants in the child condition of Gray and Wegner’s (2009) simply saw a picture of a child and read that he was 5 years of age. In our study, participants in this condition read, “Samantha is a five-year-old girl who lives with her parent and older sister Jennifer.” Thus, the difference in our methodologies alone does not explain the failure to replicate Gray and Wegner’s (2009) results, and suggests that moral typecasting may be a more fragile phenomenon than originally suggested.
Finally, while the order effect observed in the general condition was unexpected, it may be due to participants’ differential assumptions about the relative development of children and adults. We typically assume that adults are more developed than children, both physically and mentally. In this way, if participants rate an adult first, and then rate a child, they may feel compelled to provide slightly mitigated ratings across the board for the child, based upon the assumption that children are less developed. Conversely, if participants read about the child first, not only would the child’s ratings be made without reference to the adult (i.e., not be mitigated), but participants might also assign augmented ratings to the adult based again on the assumption that adults are more developed than children. Although further experimental work would need to be conducted to confirm this hypothesis, as a post-hoc explanation it accounts for the fact that participants who rated the child first provided higher ratings across the board for both the child and the adult.

General Discussion

Over the course of three studies, we investigated how blame attribution, and mind and moral perception function. These three phenomena are linked through their consideration of an individual’s agency, and similarly relate to a number of other comparable theories (e.g., punishment, Carlsmith et al., 2003; social cognition, Fiske et al., 2006; dehumanization, Haslam, 2006). While we began by examining these processes with regard to an individual with ASD, to achieve a stronger experimental manipulation and more stringent test of how these theories related to each other, we also considered how they applied to a child and adult woman.

In Study 1 we hypothesized that a target with ASD would receive mitigated intentionality, capacity, and blame judgments. Confirming this hypothesis, we found that a target with ASD was perceived to act with less intentionality and capacity (i.e., with less agency) than
an NT target, and that these judgments partially mediated reduced blame judgments, corroborating Malle et al.'s (2014) path model of blame. Study 2 examined competing predictions of how the dimensions of agency and experience (of the TMP), and moral agency and patiency (of the TDM) relate to each other, as well as how a target with ASD is perceived along these dimensions. According to the context hypothesis, we predicted that the relationship between agency and experience ratings would mirror that between moral agency and patiency ratings within each context condition (either specific or general). Based upon the TMP, we hypothesized that the target with ASD would receive lower agency and experience, and moral agency and patiency, ratings in the general context, evidencing dehumanization. While the target with ASD was perceived to possess less agency and moral agency in the general context, his perceived experience and moral patiency did not differ from the NT target. Based upon the TDM (which posits that moral agency and patiency are inversely related), we hypothesized that the target with ASD would be typecast as a moral patient and thus receive lower moral agency and agency, and higher moral patiency and experience, ratings in the specific context. While we again found that a target with ASD was perceived to possess lower levels of moral agency in the specific context, he was not perceived to possess different levels of agency, patiency, or experience than the NT target. Thus, the TDM hypothesis that moral agency and patiency, and agency and experience, ratings would be inversely related was not confirmed. Study 3 was designed as a replication of Study 2, with a stronger experimental manipulation of the difference between targets, and greater degree of fidelity to the methodologies of the original theories. The specific context prediction (that the targets’ moral agency and patiency, and agency and experience, scores would be inversely related) again received little support, while the general context prediction (that the girl would receive lower agency and moral agency ratings than the
woman, but that their experience and moral patiency ratings would not differ) was confirmed.

In sum, these studies generated considerable empirical support for both Malle et al.’s (2014) Path Model of Blame and Gray et al.’s (2007) TMP, while Gray et al.’s (2012) TDM failed to receive such support. Across three moral violations in Study 1, intentionality and capacity judgments at least partially mediated blame attribution, corroborating Malle et al.’s (2014) sequencing of the path model. Predictions concerning the relationship of agency and experience, and moral agency and patiency, ratings in the general context were based upon Gray et al.’s (2007) TMP. In Study 3, these predictions were fully confirmed, and although they were not confirmed in Study 2 (the target with ASD was not perceived to possess less experience or moral patiency as predicted), this appeared to be due to fault with our prediction that the stigma associated with ASD would lead to dehumanization, operationalized as lower agency and moral agency, and lower experience and moral patiency, ratings. Finally, based upon the TDM, we predicted that moral agency and patiency, and agency and experience, scores would be inversely related in the specific context. This hypothesis received little support, in either Study 2 or Study 3.

**Autism Spectrum Disorder, Disability Status, and Perceived Agency**

While PBS is intended to prevent the display of PB in the first place, when PB does arise it is responded to with deterrence-based disciplinary methods (Carr et al., 2002), which are typically used in response to unintentional norm violations (Darley & Pittman, 2004). Rather than simply assigning a punishment equal in severity to the norm violation, deterrence-based punishment is assigned in order to optimally prevent the display of similar behavior in the future. Consequently, the severity of the punishment assigned to individuals with ASD, within the PBS framework, is often mitigated as compared to the severity of punishment assigned to NT
students, in line with traditional disciplinary methods. Because punishment is positively correlated with blame (Darley & Pittman, 2004), this suggests that people with ASD may be blamed less when they commit norm violations, but the extant literature on PBS has not yet addressed whether this is the case. Study 1 demonstrates that people with ASD are indeed blamed less for norm violations than NT individuals, and future theoretical work on PBS should address how blame attribution affects the disciplinary methods of this program. Other factors, such as perceived intentionality, have been proposed to affect both blame attribution and the type of punishment assigned to individuals, thus future studies should additionally study how such factors (e.g., perceived control) may underlie both these broader judgments.

Although these studies answered a number of questions concerning how people with ASD are considered within the context of these theories, as well as the competing predictions generated by the TMP and TDM, several questions were raised during the experimental process that warrant further consideration. First, in Study 1, we hypothesized that reduced intentionality, capacity, and blame ratings for a target with ASD would arise when participants were informed about the characteristics of this disorder, and then read about situations in which the target committed moral violations putatively related to ASD. Unexpectedly, participants provided these mitigated ratings even when the target with ASD committed a moral violation putatively unrelated to this disorder. Two explanations may account for this finding.

First, at the start of the Study 1, participants were explicitly informed that individuals with ASD might be more prone to display aggressive behaviors and social deficits than NT individuals. Immediately after learning this, they read two vignettes in which a target displays PB and makes a social faux pas. The salience with which the description of ASD and consequent norm violations were linked may have led participants to discern the purpose of the study and
engage in socially desirable responding. As a result, participants may have simply assigned less blame to the target with ASD for all vignettes because they presumed that the experimenters desired them to do so, rather than actually considering whether aspects of ASD impacted the target’s actions.

Second, in Study 1 the target with ASD was perceived to act less agentically than the NT target. This effect was replicated in Study 2, with the joint finding that the perceived experience and moral patiency of these targets do not differ. Relatedly, Harris and Fiske (2006) have noted that people with disabilities are rated low on competency and high on warmth, which translates into low agency ratings and high experience ratings on the Mind Scale. Thus, in Studies 1 and 2 perceptions of a person with ASD mirrored those of people with disabilities more broadly (such that these individuals are perceived to possess less agency than nondisabled individuals). Reductions of perceived agency have also been noted to stem from consideration of a person with regard to a DSM diagnosis (within which ASD is classified; Gambrill, 2014), or with regard to medical settings more broadly (e.g., hospitals and doctor-patient interactions; Haque & Waytz, 2012). Thus, a substantial base of literature demonstrates that the fact that a person is perceived to be disabled (regardless of their disability) may be sufficient to reduce their perceived agency. One must then ask what factors drove the mitigated perceived agency of the target with ASD observed in Studies 1 and 2. While it is possible that participants directly considered how characteristics of ASD influenced the target’s intentionality capacity, it is also possible that the target with ASD’s disability status led participants to immediately assume that he possessed less agency than the NT target, which manifested as mitigated intentionality and capacity ratings.

**Differentiating the Dimensions of Mind and Moral Perception**
The results of Studies 2 and 3 call into question whether the dimensions of agency and moral agency, and experience and moral patiency, are distinct from one another. To begin, the TDM (which proposes the dimensions of moral agency and patiency) draws heavily upon the TMP (which proposes the dimensions of agency and experience). Despite these similarities, each of these theories posits that their respective dimensions relate differently to each other. The TDM states that moral agency and patiency are inversely related to each other, while the TMP states that agency and experience are independent of one another. We hypothesized that these different proposed relationships stem not from a difference in what each pair of dimensions measures, but rather from a difference in the contexts in which these theories consider targets. Specifically, the experimental evidence supporting the TMP has asked participants to consider targets in a general context, while the experimental evidence supporting the TDM has asked participants to consider targets in a specific context. Consequently, we hypothesized that in a general context agency and experience, and moral agency and patiency, would relate to each other independently (in line with the TMP), and that in the specific context pairs of dimensions would relate inversely to each other (in line with the TDM).

In the current studies, only the hypotheses based upon the TMP were confirmed. In Study 3, our hypothesis that in a general context participants would perceive that a girl possesses less agency and moral agency than a woman, but that these targets would not differ in perceived experience and moral patiency, was confirmed. While our hypothesis that a target with ASD would receive lower agency and experience, and moral agency and patiency, ratings was not confirmed in Study 2, this was likely due to our overestimation of the strength of stigma related to ASD. In line with the moral typecasting hypothesis of the TDM, we predicted that targets’ moral agency and patiency scores would be inversely related in a specific context, as would their
agency and experience scores. Yet, this inverse relationship was not observed in either Study 2 or 3. Furthermore, the relationship between moral agency and patiency, and agency and experience, ratings were very similar to each other in both context conditions of both Studies 2 and 3. Thus, while these studies generated further empirical support for the TMP, several aspects of the TDM—specifically, the moral typecasting hypothesis and the validity of the dimensions of moral agency and patiency—are called into question.

**The Moral Typecasting Hypothesis**

Of course, the present lack of evidence supporting the moral typecasting hypothesis does not allow one to conclude that this phenomenon is spurious: while we have two studies that fail to support the predictions generated by this hypothesis, Gray and Wegner (2009) conducted seven studies that revealed a variety of experimental evidence supporting this theory. Nonetheless, several theoretical questions are raised by the disagreement with the present results and Gray and Wegner’s (2009) data. First, one must note that the moral typecasting hypothesis is actually composed of two distinct hypotheses. According to the first, an individual is typecast based upon their general traits (e.g., perceptions that a child possesses less self-control and lower foresight typecasts them as a moral patient); according to the second, moral typecasting order effects can arise, such that an individual can be typecast by their actions in one moral situation for future moral situations (e.g., if a person acts as a moral agent in one situation, they will be typecast as a moral agent in future situations).

In Study 3, we found no evidence of these order effects. All participants read two moral vignettes depicting the target as either a moral agent (i.e., the target pushes glasses off a table, which shatter and cut a man’s leg) or moral patient (i.e., the target is pushed to the ground by a woman in a crowd). Half of participants read about the target as moral agent first while the other
half read about her as a moral patient first. According to the moral typecasting hypothesis, participants who read about her as a moral agent first should have typecast her as such, and consequently provided mitigated moral patiency ratings in the following vignette; participants who read about her as a moral patient first should have also typecast her as such, and consequently provided mitigated moral agency ratings in the following vignette. For example, this hypothesis predicts that if participants read about a woman getting pushed down at random in a crowd, they will typecast her as a moral patient and thus perceive that she acts with mitigated moral agency in subsequent moral situations. Neither of these predicted order effects were observed, but this may be due to the strength of the experimental manipulation. Specifically, the moral action depicted may need to be severe enough to allow participants to infer the target’s fundamental moral character in order for such order effects to arise.

A further consideration of the Gray and Wegner’s (2009) methodology sheds light on both the present failure to observe moral typecasting order effects as well as potential factors that may drive these order effects, if they indeed exist. First, although these authors conducted numerous studies, only one study was dedicated to demonstrating moral typecasting order effects. In this study, a target was described as sitting in a board meeting and either 1) taking the lead on a project to increase company profits at the expense of increased polluting or 2) walking out of the meeting due to his moral disagreement with such a plan. Gray and Wegner (2009) found if participants read about the target committing either of these agentic actions, they provided mitigated perceptions of the target’s patiency. Importantly, each of these actions communicate a great deal about the target’s moral character: the target’s decision to support or oppose this plan to harm the environment in exchange for increased for profits holds great moral weight and will likely lead to significant repercussions, for both the target himself as well as the
civilians who would be affected by increased pollution. Thus, it is likely that an individual would not make this decision impulsively, and that his decision would reflect his underlying moral character. As a result, participants may have perceived that the target’s actions in this situation were indicative of his fundamental moral character, and based upon this knowledge made other inferences about his moral character. For example, individuals may assume that people who are willing to stand by their convictions despite opposition (be it backlash from board members or public opinion) are less affected by aversive stimuli (be it public opposition or physical pain). In this way, the target’s actions in this vignette may have revealed enough about his moral character that participants were able to typecast him as a moral agent, and in turn provide mitigated perceptions of his moral patiency.

In contrast to this, learning of a target pushing glasses off a table or being pushed down in a crowd – as participants read in this study – provides much less information about the target’s moral character. For example, pushing a tray of glasses of a table is a relatively minor moral violation, and may be judged as an impulsive action, rather than reflective of the target’s moral character, by participants. Thus, if moral actions indeed lead to typecasting by allowing people to infer an individual’s fundamental moral character, one would not expect to observe a typecasting effect in response to these vignettes.

Returning to a comparison of the TMP and TDM more broadly, it remains unclear whether the dimensions of agency and moral agency, and experience and moral patiency, are unique from each other. At the level of operationalization, the dimensions of moral agency and patiency are subsumed within the broader dimensions of dimensions of agency and experience. Gray and Wegner (2009) operationalize moral agency as perceived intentionality and moral responsibility, which roughly map onto the “planning” and “morality” components of the seven-
item agency dimension of the Mind Scale. Patiency is operationalized simply as one’s capacity to experience pain or pleasure, both of which are present on the 11-item experience dimension of the Mind Scale. Disregarding minor operational differences in the measures used for these dimensions, the measures used for moral agency and moral patiency are almost entirely subsumed by the measures used for agency and experience.

The only novel evidence that Gray et al. (2012) present to support their claim that the dimensions of moral agency and patiency relate to each other differently than the dimensions of agency and experience stems from the experimental work of Gray and Wegner (2009) on moral typecasting. This evidence, though, has been called into question by the present studies. Furthermore, when the authors cite past literature in support of the TDM, they typically apply their theory post hoc to past experimental findings. Thus, while their theory appears to explain the findings of several studies, it cannot be confirmed as the only explanation. For example, victims are more likely to engage in blameworthy behavior than individuals who have not been victimized (Zitek, Jordan, Monin, & Leach, 2012). Gray et al. (2012) posit that this is due to victim’s belief that their behavior will be perceived as less blameworthy because they have been typecast as moral patients. It is equally possible, though, that victims are simply enraged by their situation and choose “act out,” or even that they feel entitled to commit blameworthy behavior based upon their victimization. While the moral typecasting hypothesis may explain the results of Zitek et al.’s (2012) study, so do a number of other explanations.

Finally, one must also address paradoxical situations that arise for the TDM, such as the victim-blaming that often occurs with rape victim/survivors. Because rape victim/survivors experience tremendous trauma, the TDM predicts that they would be typecast as moral patients, thus reducing their perceived levels of moral agency. Yet, we see that, contrary to this prediction,
rape victims are often blamed for their assault (see meta-analysis by Suarez & Gadalla, 2010). Gray et al. (2012) write off such evidence as invalid. They argue that studies demonstrating this victim-blaming phenomenon lead participants to feel “complicit” in the violence taking place, which then leads participants to engage victim blaming as a way to justify their own immoral behavior and alleviate their guilt (see: the Lerner paradigm; Cialdini, Kendrick, & Hoerig, 1976). This is most often not the case. For example, Bieneck and Krahé (2011), who recently further validated the victim blaming effect, collected their data by administering surveys to a college psychology class – students had no reason to believe that their responses would in any way contribute to the suffering of an actual person. Thus, nothing about the experimental design should have led participants to feel complicit in the fictitious rapes presented. Despite this lack of complicity, we see that rape victims are still being blamed for their situations, contrary to the prediction of the TDM.

Future Directions

The results of these studies have generated a number of new research questions and avenues for future research. First, the target with ASD was perceived to act with less intentionality and capacity, and was attributed less blame, than the NT target even when the moral violation was putatively unrelated to the characteristics of ASD. Two possible explanations may account for this: 1) that participants were engaging in socially desirable responding and 2) that participants perceived that the target with ASD possessed less agency based on his disability status, rather than the based upon a consideration of how the characteristics of ASD actually affected his actions. Both of these factors can be controlled for by modifying the experimental design so that participants only learn that the target has ASD, without learning about any of the characteristics of the disorder. This modification should reduce
the salient link between the characteristics of ASD and moral violations committed. Consequently, if participants were engaging in socially desirable responding, the perceived agency of the target with ASD and NT target should no longer differ, as the salient link between the target description and moral violations has been eliminated. If participants perceived that the target with ASD possessed mitigated agency based upon his disability status alone, then this target should continue to receive mitigated agency ratings with this modification.

Future studies can additionally examine how disability status affects agency perceptions more broadly. While Harris and Fiske (2006) suggest that any disability status leads to mitigated perceived agency (or competence), Study 1 calls this into question. Specifically, T1D is also a disability—such that it can cause significant impairments to daily functioning—yet the target with ASD was still perceived to act with less intentionality and capacity, and receive less blame, in all vignettes than the target with T1D. One must then ask whether the type of disability that one possesses moderates the effect that one’s disability status has on their perceived agency. Specifically, do mental disabilities affect perceived agency differently than physical disabilities?

Agency is *founded* in mental processes (principally intentionality and forethought; Bandura, 2001), while physical action is a way of *enacting* agentic desire. In this way, mental disabilities may be perceived to influence an individual’s fundamental capacity to be agentic, whereas physical disabilities may be perceived to only inhibit certain forms of agentic action. For example, if individuals learn that a person with ASD acts with mitigated intentionality in certain situations, they may generalize this recognition so that they perceive that a person with ASD has a lesser overall capacity to act intentionally, and in turn lesser capacity to act agentially. In contrast to this, T1D does not affect a person’s mental faculties, nor does it even affect their ability to enact agentic desire through physical actions, and thus one would not
expect individuals to perceive that a person with T1D possesses mitigated agency. To further assess how physical disabilities may affect perceived agency, one can consider a woman who has lost control of both of her legs. This woman still possesses the same mental faculties to develop agentic desire as an NT individual, but she has a reduced capacity to execute *certain* desires. While she cannot simply will herself to walk, she can still engage in a variety of other forms of agentic action (e.g., writing a letter, planning a vacation, or using a wheelchair to get from one place to another). Thus, physical disabilities appear to restrict agentic action in a limited and identifiable way, whereas mental disabilities may be perceived to affect a person’s fundamental capacity to develop agentic thoughts and desires. Future studies should assess whether this is indeed the case, and further tease apart the nuanced ways in which one’s disability status may affect their perceived agency.

To better understand how typecasting order effects function, future studies should manipulate the severity of the moral action. If typecasting order effects only arise when participants perceive that a moral action is indicative of a target’s fundamental moral character, then order effects should become more pronounced as the severity of the action is increased. Furthermore, future studies can directly address whether perceptions that an action is indicative of a target’s moral character mediates order effects. By asking participants the extent to which they believe each action is in line with the target’s moral character, researches can test whether this belief moderates the presence and/or strength of order effects.

Finally, further studies need to be conducted to determine whether the dimensions of moral agency and patiency are indeed unique from the dimensions of agency and experience; specifically, whether the dimensions of moral agency and patiency indeed relate differently to each other than the dimensions of agency and experience. The only experimental evidence
supporting the proposed inverse relationship between moral agency and patiency stems from Gray and Wegner’s (2009) study, yet results of the present studies contradict this evidence. Thus, future studies must begin by continuing to test the predictions generated by the TMP and the TDM with a variety of target types and moral violations in order to generate a larger base of experimental evidence. If these dimensional pairs are determined to be unique, and the hypothesized relationship between the dimensions of each pair (i.e., that the pairs are either independently or inversely related) is observed, future studies should also test whether the context moderates these relationships.

Several prominent theories of social perception and regulation (e.g., punishment: Darley & Pittman, 2003; social cognition: Fiske et al., 2006; mind perception: Gray et al., 2007; moral perception; Gray et al., 2012; dehumanization: Haslam, 2006; blame: Malle et al., 2014) overlap in their consideration of a person’s agency, including the extent to which an individual displays self-control, planning, forethought, and intentionality. Yet, agency is hypothesized to operate slightly different within each of these theories. For example, while the TMP categorizes all aspects of agentic action on their agency dimension, Haslam’s (2006) model of dehumanization includes perceived agency as a component of his HN dimension while self-control is a component of his HU dimension. Fiske et al.’s (2006) model of social cognition includes perceived intent along their warmth dimension, while perceived skill, ability, and efficacy are listed along their competence dimension. Furthermore, the TMP and TDM each propose a different relationship between their respective dimensions of agency and experience. Thus, while the literature has outlined what factors comprise agency (e.g., Bandura, 2001), it remains unclear exactly how perceived agency affects social perception. It is particularly perplexing that the components of agentic action are split unevenly between the parallel dimensions of several
theories of social perception. To obtain a more complete understanding of how perceived agency affects social perception, future studies should reexamine these theories with a specific focus on how agency—and the components that comprise agentic action—function. Future studies should also more directly interrogate the differences between current, dual-dimensional theories of social perception in order to determine whether they are indeed parallel theories measuring unique phenomena, or if they are rather only nominally different.
References


Cameron, C. D., Harris, L. T., & Payne, B. K. (2016). The emotional cost of humanity:
Anticipated exhaustion motivates dehumanization of stigmatized targets. *Social Psychological and Personality Science, 7*(2), 105-112.


Figure 1. Mean blame, perceived intentionality, and perceived capacity ratings provided by participants in the ASD and T1D conditions for the social norm violation vignette. All differences were significant at $p < .001$. 
Figure 2. Mean blame, perceived intentionality, and perceived capacity ratings provided by participants in the ASD and T1D conditions for the problem behavior vignette. All differences were significant at $p < .001$. 
Figure 3. Mean blame, perceived intentionality, and perceived capacity ratings provided by participants in the ASD and T1D conditions for the control vignette. All differences were significant at $p < .001$. 
Figure 4. Mind and moral perception by target type in the general context condition (Study 2)

*Differences significant at $p < .01$
Figure 5. Mind and moral perception by target type in the specific context condition (Study 2)

*Difference significant at $p = .01$
Figure 6. Mind and moral perception by target type in the general context condition
(Study 3)

*Differences significant at $p < .001$
Figure 7. Mind and moral perception by target type in the specific context condition (Study 3)

*Difference significant at $p < .001$
Appendix A

Neurotypical target:

Tim is a student in seventh grade. He’s pretty happy with all of his classes, and although he doesn’t care too much for English, he loves science, and learning about the world around him. He tries to do one extracurricular activity each quarter in addition to school, and he currently plays soccer. During his free time, he loves playing video games, reading books, and going on walks with his dog.

Tim also has Type 1 diabetes (T1D), which is an autoimmune disease that he was born with. T1D causes the body to destroy the cells that produce insulin, which is a hormone that enables people to get energy from food. Because there is no cure for T1D, individuals must monitor their blood glucose levels throughout the day, and give themselves insulin injections as needed to keep their levels in check. Despite this monitoring, people’s glucose levels can still become too low or too high, which is life threatening. Over time, the disease can also have serious side effects, such as kidney failure, blindness heart attack, and stroke.

Target with ASD:

Tim is a student in seventh grade. He’s pretty happy with all of his classes, and although he doesn’t care too much for English, he loves science, and learning about the world around him. He tries to do one extracurricular activity each quarter in addition to school, and he currently plays soccer. During his free time, he loves playing video games, reading books, and going on walks with his dog.

Tim also has autism spectrum disorder (ASD), a disorder that affects people in two main ways. First, individuals with ASD may be more dependent on routines, sensitive to change, and
have more repetitive behaviors or interests than typically developing people. This sometimes leads them to display aggressive or destructive behaviors when they become frustrated or angry. Second, individuals with ASD may have social deficits, which are often related to problems with their “theory of mind”, or the ability to imagine someone else’s point of view or empathize with them. As a result of these deficits, individuals with ASD may make social faux pas or say something offensive to others.
Appendix B

Social norm violation:

Tim and Alice are in the same art class. One day, Alice tells Tim that she entered her best painting of the year in the school-wide art competition, and that she really hopes to win the competition. Next week, the school announces that someone else won the competition. In art class that day, a classmate sitting at Tim and Alice’s table asks Alice if she is sad that she lost the competition. Alice had been very quiet all day, but she smiles weakly and tells the classmate that she doesn’t mind that she lost. After that, Tim talks about how he had entered the competition last year, and even though he didn’t like the piece he entered, he still won. When she hears this, Alice starts crying and leaves the classroom.

Problem behavior:

Science is Tim’s favorite class of the day. One day, though, when he arrives at his science class he finds the teacher directing students to go to the auditorium. Instead of having class, a last minute school-wide assembly has been called. Tim tells the teacher that it is unfair to cancel class with no warning, and he refuses to go to the assembly. The teacher tells Tim that he must go; otherwise he will receive a detention. After hearing this, Tim becomes upset, and yells at the teacher, saying that she is being mean. Then, he pushes over a desk and leaves the room.

Non-ASD related norm violation

Tim and Cory are both defenders on the soccer team at school. Last year, Tim had been one of the best defenders on the team, and he played every game, but Cory practiced every day over the summer, and he improved so much that this year the coach has started putting him in instead of Tim. As the season goes on, Tim becomes increasingly frustrated at having to sit on the bench while Cory plays, so he makes a plan to get more playing time. All players must wear
cleats, shin guards, and their team uniform in order to play, so Tim decides to hide all of Cory’s gear before the next game. On the day of the game, Tim moves Cory’s gear to a different locker. As a result, Cory is benched and Tim gets to play the entire game.
Appendix C

Social norm violation questions:

1) How much blame does Tim deserve for hurting Alice’s feelings?
   a. 1 (No blame at all) 2 3 4 5 6 7 (A great deal of blame)

2) Did Tim intentionally hurt Alice’s feelings?
   a. 1 (Not at all) 2 3 4 5 6 7 (Very much so)

3) Did Tim know that what he said was going to hurt Alice’s feelings?
   a. 1 (No, definitely not) 2 3 4 5 6 7 (Yes, definitely)

Problem behavior questions:

1) How much blame does Tim deserve for misbehaving?
   a. 1 (No blame at all) 2 3 4 5 6 7 (A great deal of blame)

2) Did Tim intentionally misbehave?
   a. 1 (Not at all) 2 3 4 5 6 7 (Very much so)

3) Was Tim in control of his behavior?
   a. 1 (Not at all) 2 3 4 5 6 7 (Very much so)

Non-ASD related norm violation

1) How much blame does Tim deserve for misbehaving?
   a. 1 (No blame at all) 2 3 4 5 6 7 (A great deal of blame)

2) Did Tim intentionally misbehave?
   a. 1 (Not at all) 2 3 4 5 6 7 (Very much so)

3) Was Tim in control of his behavior?
   a. 1 (Not at all) 2 3 4 5 6 7 (Very much so)
Appendix D

Dehumanization questionnaire:

All questions were answered on the following likert scale: 1 (Not at all) 2 3 4 5 6 7 (Very much so)

To what extent do you agree with the following statements?

1) Tim is broad-minded [HU]
2) Tim is fun-loving [HN]
3) Tim is impatient [HN]
4) Tim is thorough [HU]
5) Tim is disorganized [HU]
6) Tim is shy [HN]
7) Tim is active [HN]
8) Tim is stingy [HU]
Appendix E

Mind Scale Questionnaire (Gray et al., 2007):

Now, we would like your impressions of Tim. You will be asked to rate whether Tim possesses a variety of traits, by comparison to the average person. Please give us your immediate impression and be as honest as possible. All questions were answered on the following likert scale: 1 (not at all) 2 3 4 5 6 7 (very much so)

Agency
1. Self-control
2. Morality
3. Emotion Recognition
4. Memory
5. Planning
6. Communication
7. Thought

Experience
1. Hunger
2. Fear
3. Pain
4. Pleasure
5. Rage
6. Desire
7. Personality
8. Consciousness
9. Pride
10. Embarrassment
11. Joy
Appendix F

Moral Questionnaire
Now, we would like your impressions of Tim. Please consider Tim in comparison to the average person while answering the following questions.

**Compared to the average person,** to what extent can Tim…

1) Be fully morally responsible for his actions  
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)

2) Deserve blame for acting negatively?  
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)

3) Deserve praise for acting positively?  
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)

4) Experience suffering caused by another person  
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)

5) Experience pleasure caused by another person  
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)
Appendix G

Order 1:

1. Imagine that Sharon/Samantha is out to lunch. She pushes a tray of glasses off a table. They shatter and one of the shards cuts the man, Roger, sitting next to her.

2. Imagine that a woman, Tasha, is in a rush. She pushes through a crowd and knocks over a Sharon/Samantha, who falls and scrapes her knee.

Order 2:

1. Imagine that a woman, Tasha, is in a rush. She pushes through a crowd and knocks over a Sharon/Samantha, who falls and scrapes her knee.

2. Imagine that Sharon/Samantha is out to lunch. She pushes a tray of glasses off a table. They shatter and one of the shards cuts the man, Roger, sitting next to her.
Appendix H

Moral Questionnaire (Specific)
Now, we would like your impressions of what happened. *For every vignette, the first two questions inquire about the person who was portrayed as a moral agent (Person 1), while the third question inquires about the person who was portrayed as the moral patient (Person 2).*

a. How responsible is Person 1 for her behavior?
b. How intentional was Person 1’s behavior?
c. How much pain does Person 2 feel when she scrapes her knee?

Moral Questionnaire (General)
Now, we would like your impressions of Sharon/Samantha. Please consider Sharon/Samantha in comparison to the average person while answering the following questions.

**Compared to the average person,** to what extent can Sharon/Samantha…

6) Be responsible for her actions
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)

7) Act intentionally?
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)

8) Experience pain caused by another person
   a. 1 (much less so) 2 3 4 5 6 7 (much more so)