Thinking About Other Minds

By

Christopher Brett Jaeger

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

January 31, 2020

Nashville, Tennessee

Approved:

Daniel T. Levin, Ph.D.

Sarah Brown-Schmidt, Ph.D.

Jonathan D. Lane, Ph.D.

Owen D. Jones, JD

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Table Page

INTRODUCTION

People engage in countless interactions every day. Often, these interactions are face-to-face, real-time encounters with other people. We have conversations with friends, play in soccer matches, or attend classes. Navigating these interactions requires us to think about others' minds. We use our observations to construct the current beliefs, desires, and goals of the people with whom we are interacting, and shape our own behavior accordingly. Psychologists call this "theory of mind" (e.g. Apperly, 2012).

Importantly, however, we regularly think about others' minds in a variety of *other* ways. We think about minds of people we cannot readily observe—for example, when driving (does that driver in the other lane want to merge?). We think about what people's mental states were in the past—for example, when serving as a juror in a car accident case (was the driver paying attention?). We think about aspects of mind beyond beliefs, desires, and goals—for example, when evaluating others' capacities (are they intelligent?). And sometimes, we think about "minds" that do not belong to people at all—for example, minds of pets, superheroes, or God (Epley, Akalis, Waytz, & Cacioppo, 2008; Lane, Wellman, & Evans, 2010; Lane, Evans, Brink, & Wellman, 2016). All of these examples clearly involve thinking about other minds, but they fall outside the scope of "theory of mind," at least as traditionally conceived.

Theory of mind is generally defined as the capacity to mentally represent other people's mental states (e.g. beliefs, desires, goals) as distinct from one's own, and to use those representations to explain or predict others' behavior (e.g., Premack & Woodruff, 1978; Baron-Cohen, Leslie, & Frith, 1985; Westra, 2017). But while most researchers would likely agree with a definition in this vein, "the appearance of consensus on what theory of mind is, and how we should study it, is misleading" (Apperly, 2012). Closer examination reveals substantial

disagreement with respect to the nature of theory of mind, the body of cognitive skills it entails, and when and how people use it.

The diversity of views of theory of mind is illustrated by the diversity of measures researchers use to assess it. The classic theory of mind task is the "false-belief" task, typically conducted with children (Wimmer & Perner, 1983). Participants view a skit or movie in which one agent sees an object (e.g. a marble) placed in one location, and then exits the scene. The participant then sees the location of the object change in the agent's absence. The agent returns, and the participant is asked where the agent will look for the object. Through this task, "theory of mind" is operationalized as the ability to track the agent's *beliefs* about the state of the world and use those beliefs to predict the agent's behavior.

In recent years, however, research concerning theory of mind has expanded far beyond the province of false beliefs. For example, researchers have developed tasks probing how readily participants infer motives underlying statements and behaviors (the "strange stories" and "silent films" tasks, see Lecce, Bianco, Devine, & Hughes, 2017), and read emotions from others' facial expressions (the "mind in the eyes" task, Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). Other researchers have come to view visual perspective taking in both children and adults as a skill related to the deeper cognitive inferences that drive theory of mind (e.g. Apperly, 2012; Elekes, Varga, & Király, 2016), and have adapted perspective taking tasks to study those inferences (e.g., the "dot task" (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010) and the "number task" (Surtees, Butterfill, & Apperly, 2012)). While the various tasks researchers are using to study theory of mind may seem to tap distinct cognitive skills, they tend to be strongly correlated and load onto a common latent variable in structural models (see Bigelow & Dugas, 2009; Jones et al., 2018; Devine, White, Ensor, & Hughes, 2016). This

broadening understanding of how people think about others' mental states has corresponded with shifting terminology; many researchers now speak in terms of "mindreading" (Apperly, 2010; Carruthers, 2016) or "perspective taking" (e.g. Surtees, Butterfill, & Apperly, 2012) rather than, or in addition to, "theory of mind."

While scientific understanding of theory of mind (or mindreading, or perspective taking) is broadening, a number of ways in which people think about other minds still lie outside its scope, as suggested at the outset of this introduction. What inferences do people draw about others' minds when the others are not specific, observable individuals? Or are not people at all? What inferences do people make about others' capacities? What cognitive processes are deployed in making these inferences?

This dissertation explores these questions in three different contexts. Chapters 1 and 2 of this dissertation focus on two relevant legal settings. Law affords a particularly interesting avenue for investigating these questions, as law routinely asks people to reason about unseen causes of events.

Chapter 1 examines how people think about law's reasonable person standard. The legal system often requires decisionmakers to compare a defendant's conduct to that of an abstract "reasonable person." On one view, at least, this involves thinking about others' minds—decisionmakers must imagine a "reasonable person," attribute mental characteristics to that person, and then draw inferences about how that person would behave in a particular set of factual circumstances. In doing so, people might base their mentalistic attributions on their specific observations about themselves and others (see Jaeger, Levin, & Porter, 2017; Feigenson, 1995), or leverage broader concepts of a "generalized other" (see Mead, 1934). However, some legal scholars have argued that applying the reasonable person standard can—and should—use

strategies that do not require detailed mentalistic reasoning at all (Miller & Perry, 2012). The series of studies reported in Chapter 1 experimentally probe how decision makers think about what a reasonable person would do in a legally-relevant situation. The results suggest that decision makers attempt to mentally simulate what an average person would do under the relevant circumstances, raising the possibility that systematic biases in our understandings of the human mind may lead to systematically biased decisions in some types of legal cases.

While Chapter 1 concerns how theory of mind is applied to an abstract, imagined human, Chapter 2 focuses on ways in which theory of mind may be applied to *non-humans*. People are so accustomed to (and adept at) using theory of mind to explain the world around them that they sometimes *over*use it. Anthropomorphism occurs when people apply theory-of-mind-type reasoning to entities and objects that are not human (Epley, Waytz, & Cacioppo, 2007). The human tendency to anthropomorphize is enhanced when people encounter the unfamiliar and unpredictable; people are especially likely to attribute human-like mental states to technological gadgets (ibid.; Epley, Akalis, Ways, & Cacioppo, 2008). This tendency may be relevant to contemporary issues of law and policy, as our society prepares to deal with an influx of autonomous devices such as self-driving cars (Jaeger & Levin, 2016). Chapter 2 presents three experiments investigating the cognitive processes involved when people attribute human-like mental qualities to self-driving cars, and how those processes shape legal decisions and policy opinions about such cars. The results suggest that anthropomorphism is multidimensional: some, but not all, types of human thought are likely to be attributed to machines, and the different types of attributions have distinct (and sometimes opposing) effects on their legal judgments and policy opinions.

Finally, Chapter 3 of this dissertation explores people's puzzling inclination to spontaneously take the visual perspective of pictorial representations of agents, even when there is no communicative value in doing so. While this phenomenon has been documented previously (Tversky & Hard, 2009; Todd et al., 2015), little work has investigated the factors that facilitate and suppress this form of perspective taking. Chapter 3 presents five experiments examining the extent to which people shown a picture of a schematic face will spontaneously adopt its perspective, and investigates the role of context and capacity in facilitating or suppressing such perspective taking. With respect to context, I find that cues designed to prompt consideration of a schematic face as an agent facilitate perspective taking. With respect to capacity, I report mixed results about whether cognitive load suppresses perspective taking.

The studies reported in this dissertation continue a trend in the literature of broadening inquiry into how we conceptualize other minds. Most research on theory of mind focuses on online inferences of others' beliefs, desires, and goals, in support of face-to-face social interactions with other people. This dissertation explores thinking about other minds in different situations and along different dimensions. By exploring how people represent the mental states and capacities of an abstract reasonable person, a technological agent, and a schematic representation of an agent, I seek to understand further the broader question at the heart of social cognition: how do we think about other minds?

CHAPTER 1

CONSTRUCTING AN ABSTRACT MIND: THE EMPIRICAL REASONABLE PERSON

Abstract

In American tort litigation, decision makers evaluate whether a litigant's actions were negligent by comparing the litigant's actions to those of a "reasonable person." Applying this standard poses a substantial cognitive challenge, as decision makers must assign characteristics to the abstract reasonable person and use them to draw inferences about behavior. But what characteristics do decision makers assign? This Chapter examines three possibilities discussed in the legal literature. First, the reasonable person might be defined based on intuitions about what a roughly average person would do (the "empirical reasonable person"). Second, the reasonable person might be defined based on intuitions about what the very best and most capable among us would do (the "aspirational reasonable person"). Third, the reasonable person might be defined purely by normative principles like utility maximization, regardless of how actual people behave (the "economic reasonable person"). This Chapter describes four experiments that probe whether lay decision makers tend to define the reasonable person in empirical, aspirational, or economic terms. The results suggest that, of these possibilities, lay decision makers generally favor the empirical reasonable person standard. This finding lays a foundation for future work bringing psychological research on people's understanding of others' minds to bear on legal scholarship. If legal decision makers endeavor to deploy an empirical reasonable person standard, but systematically misjudge others' cognitive capacities (as prior psychological research suggests), then verdicts may be systematically biased.

I. INTRODUCTION

The reasonable person standard lies at the center of our tort system, separating negligent conduct (for which actors bear the cost) from faultless conduct (for which the costs lie where they fall). But implementing the standard is cognitively challenging. Factfinders must imagine the circumstances of a particular case and determine how a hypothetical reasonable person would have acted under those circumstances. To make this exercise tractable, factfinders must assign cognitive characteristics to the reasonable person.[1] But what characteristics do they assign? And what principles guide their assignment?

These questions have been the subject of much debate in the legal literature. Some scholars argue—or assume—that the reasonable person is defined based on the decision makers' intuitions about what a roughly average person would do in the circumstances (the "empirical reasonable person standard"). Others contend the reasonable person is defined based on the decision makers' intuitions about what the very best and most capable among us would do in the circumstances (the "aspirational reasonable person standard"). I refer to the empirical and aspirational hypotheses collectively as "positive hypotheses," as both are grounded in beliefs about actual human behavior. In contrast to these positive hypotheses, other scholars argue that the reasonable person is, or at least should be, defined by universal logical or ethical principles, regardless of how actual people behave. Most commonly, these scholars work from rational actor models popular in economics, contending that the reasonable person always chooses the utility-maximizing course of conduct (the "economic reasonable person standard").[2]

---

[1] Jeffrey J. Rachlinski, *Misunderstanding Ability, Misallocating Responsibility*, 68 BROOKLYN L. REV. 1055 (2003).

[2] *See* Alan D. Miller & Ronen Perry, *The Reasonable Person*, 87 N.Y.U. L. REV. 323 (2012).

Despite ongoing debate surrounding these possibilities, no research to date has empirically examined how people construe the reasonable person in practice. Scholars have told us much about who the reasonable person should be, but we know shockingly little about who the reasonable person is. In this Chapter, I present a series of four experiments investigating lay constructions of the reasonable person. These experiments provide an initial test of whether lay decision makers tend to define the reasonable person in empirical, aspirational, or economic terms, and examine whether the answer applies across the board or depends on the cognitive characteristics at issue. My findings provide what is, to my knowledge, the first experimental evidence that lay decision makers' interpretations most align with the empirical reasonable person standard, and that this is the case across a variety of cognitive characteristics.

This Chapter proceeds as follows. Part II situates the reasonable person standard in American tort law and reviews the prominent theoretical positions on how the reasonable person should be defined. Part III presents a series of four original experiment investigating whether lay decision makers tend to favor an empirical, aspirational, or economic definition. In each experiment, participants played the part of jurors, deciding a series of negligence cases described by written case vignettes. Each case vignette included information that was critical under at least one definition of the reasonable person: information about what portion of the population would have acted differently under the circumstances (critical on an empirical or aspirational view), or about whether the defendant's course of conduct was utility-maximizing (critical on an economic view). Observing what type(s) of information affected participants' negligence judgments allowed me insight into how participants conceptualize the reasonable person. Part IV discussed my findings and their implications for tort law. Part IV also outlines directions for future research.

II. TORT LAW'S REASONABLE PERSON

People get injured. Often, they get injured as a consequence of others' actions. But who bears the cost of such injuries? In the United States, the default rule has long been that those who suffer injuries bear their own costs.[3] Generally speaking, so long as people act reasonably, any harm that results from their actions will "lie where it falls."[4] Tort law, however, provides exceptions to this default rule. Tort claims offer mechanisms for people and entities who suffer injuries due to others' conduct to obtain recompense from those others, provided certain conditions are satisfied.[5] The most common condition is that the injury-causing other acted negligently—that he or she failed to act as a *reasonable person* would have acted under the circumstances.

A. The Reasonable Person as the Standard for Negligence

To prevail on a negligence claim—and thus recover compensation from the defendant—a plaintiff must sufficiently demonstrate (1) that he or she suffered an *injury*, (2) that the defendant owed a relevant *duty* of care to the plaintiff, (3) that the defendant *breached* that duty of care, and (4) that the defendant's breach was the *cause-in-fact* and *proximate cause* of the plaintiff's injury.[6]

---

[3] Rachlinski, *supra* note 1, at 1055 ("An important default principle of tort law in the Anglo-American legal tradition is that harm must 'lie where it falls.'").

[4] *See* Gideon Rosen, *Skepticism About Moral Responsibility*, 18 Philosophical Perspectives 295, 301 (2004) ("You are never obliged to take every possible step, no matter how costly, to ensure that no one is harmed by what you do. You are required only to take certain reasonable steps. If you do that much and harm results anyway, then in the vast majority of cases the harm must 'lie where it falls.'");

[5] John C.P. Goldberg, Anthony J. Sebok, & Benjamin C. Zipursky, TORT LAW: RESPONSIBILITIES AND REDRESS, 3 (2004). For this reason, some scholars have referred to tort law the law of "private and privately redressable wrongs." John C.P. Goldberg & Benjamin C. Zipursky, *Torts as Wrongs*, 88 Tex. L .Rev. 917, 918 (2010).

[6] Goldberg, Sebok, & Zipurksy, *supra* note 5.

The reasonable person standard is foundational to the duty and breach elements. With respect to duty, the general rule is that people owe one another "an unqualified duty to conduct [themselves] with reasonable care for the person and property of others."[7] A person breaches the duty of care when he or she fails to act reasonably carefully—that is, when he or she fails to act as carefully as the ordinary, reasonable person would have acted under the circumstances.

Through its role in defining the duty and breach elements, the reasonable person standard functionally divides conduct that is negligent (for which the actor will bear the cost of any resultant injuries) from conduct that is faultless (for which the actor will not bear such cost).[8] Stated differently, a person is (in theory) assured that he or she will not be liable for negligence so long as he or she acts as a reasonable person would act.[9] This reflects a recognition that it is impracticable, if not impossible, for people to take all possible precautions at all times.[10] Almost every activity that people engage in creates some risk of injury to others, but "tort law is not meant to convert everyone into insurers whenever they undertake any action."[11] "[T]he standard man is

[7] *Id.* at 51.

[8] Rachlinski, *supra* note 1, at 1055 ("Defining negligent conduct and administering this definition properly is . . . critical to determining who bears the cost of accidents" and definitions of negligence "all revolve around the reasonableness of a party's behavior.").

[9] Restatement (Second) of Torts § 283 ("the standard of conduct to . . . avoid being negligent is that of a reasonable man under like circumstances.").

[10] Rosen, *supra* note 4, at 301. It has also been stated that "[t]he history of the development of the negligence standard in the early nineteenth century suggests that the *ordinary reasonable person* standard of conduct was adopted in order to preserve the jury's historic role in judging the wrongfulness of the defendant's conduct in tort actions." Patrick J. Kelly & Laurel A. Wendt, *What Judges Tell Juries About Negligence: A Review of Pattern Jury Instructions*, 77 CHI.-KENT. L. REV. 587, 588 (2002).

[11] Rachlinski, *supra* note 1, at 1055 (citing James A. Henderson, Jr., *Expanding the Negligence Concept: Retreat from the Rule of Law*, 51 IND. L. J. 467, 524-25 (1976)).

not infallible" and "mistakes in judgment which the standard man might have made in light of [his] limitations will not amount to negligence."[12]

Importantly, the reasonable person standard is generally understood as an objective standard, rather than a subjective one. This means two things. First, it means that the inquiry is whether the defendant's (external) *conduct* was reasonably careful, not whether the defendant's (internal) *intention* was to be careful.[13] Second, it means the standard is not typically tailored to the particulars of the defendant: the defendant's conduct is compared to the reasonable conduct of a *generic person*, rather than to the reasonable conduct of a person *with the defendant's specific attributes*.[14] There are exceptions to this second form of objectiveness, based on things like age (a seven-year-old is held to the standard of the ordinary, reasonable seven-year-old),[15] expertise (a doctor is held to the standard of the ordinary, reasonable doctor),[16] and physical disabilities (a blind man is held to the standard of a reasonable blind man).[17] But a vast majority of the time, adults

[12] Fleming James, Jr., *The Qualities of the Reasonable Man in Negligence Cases*, 16 MO. L. REV. 1, 4-5 (1951).
[13] Goldberg, Sebok, & Zipurksy, *supra* note 5, at 157.
[14] *Id.*; *see also* Rachlinski, *supra* note 1, at 1065 ("The test is not whether someone felt that he did his best to avoid harm, given his own personality, concerns, and interests, but whether a reasonable person would have been able to do so."); *Vaughan v. Menlove*, 132 Eng. Rep. 490 (C. P. 1837); *Burch v. American Family Mutual Insurance Co.*, 543 N.W.2d 282 (Wis. 1996).
[15] Jean Macchiaroli Eggen & Eric J. Laury, *Toward a Neuroscience Model of Tort Law: How Functional Neuroimaging Will Transform Tort Doctrine*, 13 COLUM. SCI. & TECH. L. REV. 235, 237 (2012) ("Minors are held to a standard of care appropriate for a person of the actor's age, intelligence, and mental capacity.") (citing *Bragan ex. Rel. Bragan v. Symanzik*, 687 N.W.2d 881, 884-85 (Mich. Ct. App. 2004)).
[16] *See* W. Prosser, Law of Torts, δ 32 ("Those who undertake any work calling for special skill are required not only to exercise reasonable care in what they do, but also to possess a standard minimum of special knowledge and ability.").
[17] See RESTATEMENT (THIRD) OF TORTS: PHYS. & EMOT. HARM § 11 (a) ("The conduct of an actor with a physical disability is negligent only if the conduct does not conform to that of a reasonably careful person with the same disability." (emphasis added)).

who lack discrete physical disabilities are held to the general standard of reasonableness in the community, not an individualized standard.[18]

This point is illustrated by the famous English case of *Vaughan v. Menlove,*[19] a case often described as the origin of the reasonable person standard[20] (though the exact origin of the standard is disputed[21]). Menlove built a hay stack near the edge of his property line. Despite repeated warnings that his hay stack was a fire hazard, Melove did a poor job of shaping and maintaining the hay stack so as to minimize the risk. The stack ultimately caught fire and burned down his neighbor Vaughan's cottages. Vaughan sued. Menlove contended that he was not liable because he had built and maintained the hay stack to the best of his (poor) ability. The English court ruled that Menlove was liable: even if Menlove did his best, his actions did not satisfy the relevant standard of how a man of "ordinary prudence" would act under the circumstances. Menlove was not excused from liability because his hay-stack-building abilities were below par.

While sympathy for Menlove tends to be in short supply, the application of *Menlove*'s principle in other cases has caused much consternation. Consider *Burch v. American Family Mut. Ins. Co.*[22] In that case, fifteen year-old Amy Burch, who "was born with cerebral palsy and mental

---

[18] James, *supra* note 12, at 1-2 ("[M]any of the actor's shortcomings such as awkwardness, faulty perception, or poor judgment, are not taken into account if they fall below the general level of the community.").

[19] 132 Eng. Rep. 490 (C. P. 1837).

[20] Ashley M. Votruba, *Will the Real Reasonable Person Please Stand Up: Using Psychology to Better Understand How Juries Interpret and Apply the Reasonable Person Standard*, 45 Ariz. St. L.J. 703, 707 (2013) ("[M]any cite the 1837 case, *Vaughan v. Menlove*, as the first case mentioning the reasonable man.").

[21] Randy T. Austin, *Better Off with the Reasonable Man Dead or The Reasonable Man Did the Darndest Things*, 1992 BYU L. REV. 479, 480-81 (1992). While many (including William Prosser) identify *Vaughan v. Menlove* as the origin of the reasonable person, others (including Professor Ronald Collins) date the origin of the reasonable person standard earlier, possibly as far back as 1796. *See* Ronald K.L. Collins, *Language, History, and the Legal Process: A Profile of the "Reasonable Man"*, 8 Rut.-Cam. L.J. 311, 312 (1977).

[22] 543 N.W.2d 277 (Wis. 1996).

retardation with autistic tendencies and functions at the cognitive level of a preschooler,"[23] turned the keys of the ignition of her father's truck, pinning her father against a building and injuring him. One issue in the case was whether the objective reasonable person standard applied to Amy. Wisconsin's Supreme Court concluded that it did, "hold[ing] that generally a tortfeasor's mental capacity cannot be invoked to bar civil liability for negligence."[24] As with any objective standard, it is inevitable that some individuals will be held responsible for "failing live up to a standard which as a matter of fact they cannot meet."[25]

For better or worse, the objective reasonable person standard is entrenched in American tort law.[26] The abilities and virtues of the reasonable person against whom a negligence defendant is judged depend on the abilities and virtues expected of the community at large, and *not* on the abilities and virtues of the particular defendant. That tells us *something* about the attributes of tort law's reasonable person—but not all that much. The reasonable person remains largely an abstract "creature of the law's imagination,"[27] an empty container for finders of fact to fill with meaning in particular cases. But to render a negligence verdict, the container must be filled—with details down to the basic cognitive processes that might have enabled an ordinary reasonable person to avoid the accident the defendant caused.[28] Legal theorists have written much about how this container should be filled.

---

[23] *Id.* at 278.
[24] *Id.* at 280 ("[W]e find that the court of appeals in *Burch I* erred in concluding that the reasonable person standard did not apply to Amy. Therefore, we overrule *Burch I* and hold that generally a tortfeasor's mental capacity cannot be invoked to bar civil liability for negligence.").
[25] James, *supra* note 12, at 2.
[26] Ian J. Cosgrove, Note, *The Illusive Reasonable Person: Can Neuroscience Help the Mentally Disabled,* 91 Notre Dame L. Rev. 421, 427 (2015).
[27] F.V. HARPER & F. JAMES. THE LAW OF TORTS 902 (1956).
[28] Rachlinksi, *supra* note 1, at 1056 ("[D]etermning whether a reasonable person could have avoided an accident requires courts to endow the hypothetical reasonable person with cognitive abilities.").

B. Scholarly Debate About the Reasonable Person

Legal thinkers have long debated how the reasonable person standard *should* be defined. "The primary question has always been whether the content [of the reasonable person standard] should be normative or positive," as those terms are used in the legal scholarship. I will discuss the positive alternative first.

Positive definitions are "founded on the idea that the reasonable person's characteristics can be deduced by observation,"[29] or "approximated using empirically observable data."[30] In other words, the standard is based on the decision maker's experiences with other people, rather than on pure logic or abstract universal principles.

One approach to defining the reasonable person positively is to assume that he or she behaves how "the great mass of mankind" behaves[31]—that reasonableness is tantamount to conformance with "statistically prevalent norms of conduct."[32] This is what I refer to as the empirical reasonable person hypothesis. The idea is that, if a majority of people would not choose or perceive or remember something, the reasonable person should not be expected to either.[33] As explained by Professors Alan D. Miller and Ronen Perry:

> This idea of the reasonable person borrowed heavily from the concept of *l'homme moyen* (the average man) developed by the nineteenth-century Belgian statistician Adolphe Quetelet. Quetelet's average man was a representative person formed by averaging measurable variables such as height, weight, and propensity for criminal behavior. Quetelet's specific statistical approach was later discredited, but the idea of an average person

[29] Miller & Perry, *supra* note 2, at 327.
[30] *Id.* at 371.
[31] *Osborne v. Montgomery*, 234 N.W. 372 (Wis. 1931).
[32] Heidi M. Hurd & Michael S. Moore, *Negligence in the Air*, 3 THEORETICAL INQUIRIES L. 333, 377 (2002).
[33] Christopher Brett Jaeger, Daniel T. Levin, & Evan Porter, *Justice is (Change) Blind: Applying Research on Visual Metacognition in Legal Setting*s, 23 PSYCH. PUB. POL'Y & LAW 259 (2017).

> with whom an actual person can be compared has survived and forms the basis of the positive definition of the reasonable person.[34]

Legal scholars who favor this empirical reasonable person standard have described the reasonable person as "the man in the street," or "the man in the Clapham omnibus."[35]

While there may be some intuitive appeal to defining the reasonable person based on "the great mass of mankind," there are also drawbacks to doing so. Mankind, after all, is quite fallible. Indeed, there may be many situations in which the average person's conduct falls short of what we, as a society, want to incentivize through tort law.[36] One way to address this problem is to define the reasonable person not with reference to an average person, but rather with reference to the best, most careful, and most capable among us. This is what I refer to as the "aspirational reasonable person standard."

The aspirational reasonable person standard is a positive standard because it is based on, and constrained by, observations and beliefs about actual human behavior. But, on the aspirational view, the reasonable person is more prudent than the average person. Indeed, the aspirational reasonable person is "the embodiment of all the qualities which we demand of the good citizen ... if not exactly a model of perfection."[37] No one lives up to the standard of this idealized reasonable person all the time—but, importantly, it is *possible* for someone, somewhere to live up to the aspirational reasonable person standard in any particular situation (or, at least, the decision maker believes it is).[38]

---

[34] Miller & Perry, *supra* note 2, at 370.
[35] Austin, *supra* note 21, at 485; JOHN G. FLEMING, THE LAW OF TORTS 107 n.9 (4th ed. 1971)).
[36] Rachlinski, *supra* note 1, at 1061-63.
[37] FLEMING, *supra* note 35, at 107.
[38] *See, e.g.,* A.P. HERBERT, MISLEADING CASES IN THE COMMON LAW 12 (1930) (describing the reasonable person as "devoid, in short of any human weakness, but odious character who stands like a monument in our Courts of Justice, vainly appealing to his fellow-citizens to order their lives after his own example.").

The aspirational reasonable person is perhaps most clearly viewed through a statistical lens. Society contains a multitude of people with a multitude of different attributes. Some people are more careful than others, some more perceptive, some more attentive, and so on. For any given attribute X, one can imagine that all the members of society are arranged in a distribution from those with the most X to those with the least X. For example, assume that Loretta is society's most conscientious driver. She never speeds. She never cuts other drivers off. She would never use her cellular phone on the road—in fact, she cannot, because she always puts it in her trunk before driving, to resist temptation. At the other end of the spectrum is Conner, who is society's least conscientious driver. He speeds everywhere, he cuts other drivers off, and he is constantly texting and watching movies on his cellular phone while he drives. Everyone else in society falls somewhere on the distribution between Loretta and Conner.

If the reasonable person standard is what I have labeled an empirical standard, then the reasonable person falls right in the middle of this societal distribution, in or around the fiftieth percentile.[39] But, if the reasonable person is defined *aspirationally*, then he or she will be somewhere higher in the distribution. Where, exactly, is an open question—one that prior scholarship has not addressed with any specificity. For my purposes, I define the aspirational reasonable person as falling in or above the ninetieth percentile of the distribution.[40] Critically, however, the aspirational reasonable person cannot supplant Loretta on the high end of the distribution (i.e. in the one-hundredth percentile). Thus, the aspirational reasonable person, as I define it, is a positive standard, constrained by empirical limits on human behavior—or, at least,

[39] *See* Part II.B.1.a, *supra*.

[40] The ninetieth percentile seems like it could resonate as an aspirational threshold, as a grade of 90% is the cutoff for an "A" grade in most American schools. While a grade of 90% does not mean one performs better than 90% of the population (or vice versa), the 90 figure may provide an intuitive cutoff.

by the decision maker's understanding of those limits.[41] The aspirational reasonable person cannot do something that no human could do under the circumstances.

Aspirational interpretations of the reasonable person have sometimes been the subject of mockery.[42] Scholars have noted that the seemingly-absurd lengths to which the reasonable person sometimes goes in the name of caution, "disobey[ing] the direct requests of a gunman at point blank range" or "get[ting] out of his car at every railroad crossing to check for oncoming trains."[43] Nevertheless, an aspirational standard has advantages. It incentivizes citizens to be more careful than the average person, which is likely a good thing.[44] Further, use of an "idealized standard makes the law's inquiry intuitive and tractable."[45] In many cases, it might be easier for decision makers to imagine what people *can* do than to estimate what most people do.

While legal scholars long assumed the reasonable person was defined positively—either empirically or aspirationally—more recent scholarship at the intersection of law and economics has argued the reasonable person is, or, at least, should be, defined normatively, as legal scholars use the term. They contend that a normative definition is preferable because normative definitions flow not from limited observations of human behavior but from universal ethical principles[46]—

[41] If a decision maker concludes that the reasonable person would do something that the decision maker does not believe any actual person could do, then the decision maker is not applying an aspirational reasonable person standard but rather applying some sort of normative reasonable person standard. *See* Part II.B.2, *infra*.
[42] *Id*.
[43] Austin, *supra* note 21, at 489 (citing *Noll v. Marian*, 32 A.2d 18, 19-20 (Pa. 1943); *Baltimore & Ohio R. Co. v. Goodman*, 275 U.S. 66, 69-70 (1927).
[44] Rachlinski, *supra* note 1, at 1061-62.
[45] *Id.* at 1065.
[46] *Id.* at 327.

particularly, utilitarian principles.[47] Indeed, Professors Miller and Perry have argued that *any* workable positive definition of the reasonable person is logically impossible.[48]

The most famous utilitarian definition of the reasonable person is Judge Learned Hand's "Hand Formula."[49] Per the Hand Formula, decision makers should "measure and consider three things when determining negligent liability: (1) the size of the loss if an accident occurs, (2) the probability of that accident occurring, and (3) the [burden] of taking precautions that would prevent the accident from happening."[50] "[I]f the probability be called P; the injury, L; and the burden, B; liability depends upon whether B is less than L multiplied by P: whether B < PL."[51] I refer the view that the reasonable person is defined by the Hand Formula as the "economic reasonable person hypothesis."

On the economic reasonable person hypothesis, "[t]he standard is predetermined by a particular normative commitment—namely, cost efficiency—regardless of the prevailing perception [or conduct] in the relevant society."[52] That is, if decision makers apply an economic reasonable person standard, a defendant's conduct is negligent whenever he or she fails to take cost-effective precautions, even if there is not a single person in society who would have taken the precautions under the circumstances of the case. Conversely, "conduct is deemed reasonable if it

---

[47] While utilitarian definitions of reasonableness are most frequently discussed normative definition of reasonableness, there are others. For instance, the Kantian "equal freedom" definition proceeds from the normative premise that every person enjoys certain dignity and freedom that should not be encroached upon by others—that people are ends, not means. *Id.* at 348. Like Hand's formula, this definition has occasionally appeared in the language of common law tort decisions. *See id.* at 351, n. 126 (reviewing cases).
[48] Miller & Perry, *supra* note 2, at 326.
[49] *United States v. Carroll Towing Co.*, 159 F.2d 169 (2d Cir. 1947).
[50] Votruba, *supra* note 20, at 711.
[51] *Carroll Towing*, 159 F.2d at 173.
[52] Miller & Perry, *supra* note 2, at 326.

is cost-effective, even if no one truly believes it to be reasonable."[53] On the economic reasonable person hypothesis, decision makers need not be particularly concerned with empirical observations or intuitions about what people actually do in circumstances similar to those at issue in the case.

The economic reasonable person hypothesis has been tremendously influential. It has been heavily relied upon in legal scholarship, and the American Law Institute largely adopted the Hand Formula as the standard of reasonableness in the *Restatement (Third) of Torts*:

> A person acts negligently if the person does not exercise reasonable care under all the circumstances. Primary factors to consider in ascertaining whether the person's conduct lacks reasonable care are the foreseeable likelihood that the person's conduct will result in harm, the foreseeable severity of any harm that may ensue, and the burden of precautions to eliminate or reduce the risk of harm.[54]

However, some scholars have criticized the *Restatement (Third) of Torts*' inclusion of the Hand formula on grounds that it "is rarely cited and seldom applied by American courts."[55]

C. A Lay Perspective of the Reasonable Person

Debate about how the reasonable person *should* be defined is both interesting and important, but it has relatively little bearing on how lay decision makers encounter, understand, and apply the standard. The juror's experience with the reasonable person standard is much different than that of legal theorists, typically coming at the end of the exhausting effort of sitting through a trial.[56] "Amid the highly technical, jargon-filled instructions," the juror notices the judge

---

[53] *Id.*

[54] RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 3.

[55] Miller & Perry, *supra* note 2 at 333; Steven Hetcher, *Non-Utilitarian Negligence Norms and the Reasonable Person Standard*, 54 VAND. L. REV. 863, 864-65 (2001).

[56] The juror's experience is aptly summarized by Professor Ashley Votruba, *supra* note 20, at 704-06.

saying something about reasonable people exercising "ordinary care."[57] The juror likely gets no more guidance than this about the reasonable person standard.[58] The judge may not be even permitted to answer clarifying questions about the instruction.[59] Jurors are generally not told whether or how to use their empirical observations, the Hand formula, or community norms to reach a decision. "Instead they are largely left to their own devices to decide what is considered negligent behavior in this circumstance with only the vague, undefined concept of the reasonable person as their guide."[60]

So what do jurors do? "Although scholars and legal theorists have spent much time discussing how the reasonable person standard should be understood, conceptualized, and modified, little attention has been paid to how jurors actually interpret and apply the standard as presented by jury instruction"—despite the fact it is "juries who apply the standard, not legal scholars."[61] This Chapter begins to address this gap in the literature, empirically examining whether mock jurors use an empirical, aspirational, or economic definition of the reasonable person to determine liability in negligence cases.[62]

In considering how jurors define the reasonable person, a useful perspective might be found in the work of early-20th-century philosopher and social scientist George Herbert Mead. Mead posited that, in order to develop a meaningful sense of self, humans must develop and deploy the

[57] Votruba, *supra* note 20, at 704.
[58] For a survey of jury instructions concerning the reasonable person, *see* Kelly & Wendt, *supra* note 10.
[59] Votruba, *supra* note 20, at 706.
[60] *Id.*
[61] *Id.*
[62] Rachlinski, *supra* note 1, at 1056 ("[D]etermning whether a reasonable person could have avoided an accident requires courts to endow the hypothetical reasonable person with cognitive abilities.").

perspective of a "generalized other"—a broader, intersubjective viewpoint on individual actions.[63] The idea is that, through repeated social experiences, humans develop not only the ability to take the perspective of specific others with whom they are interacting, "but to generalize such particulars to members of the social groups in which we grow and develop in general."[64] Thus, the "generalized other" is a product of experience—"a composite representative of others, of society, within the individual,"[65] somewhat akin to Sigmund Freud's conception of the superego. The ability to adopt the perspective of a "generalized other" allows us to evaluate our own behavior in terms of what is *expected* of us.[66]

The "generalized other" is often illustrated through the example of team sports.[67] Consider a child learning to play soccer. Initially, the child learns by kicking a ball back and forth in a dyad, with a parent. In doing so, the child learns about the particular response patterns of his or her parent, and may account for the (specific) perspective of the parent. However, when the child begins playing in organized soccer games, he or she must learn the behavioral patterns associated with every position on the field, within the rules and context of the game. With experience, the child will internalize these patterns, and "come to 'view' [his or her] own behaviors from the perspective of the game as a whole, which is a system of organized actions."[68] In Mead's view, the same general process holds for society more broadly. For example, according to Mead, a

---

[63] George Herbert Mead, MIND, SELF, AND SOCIETY FROM THE STANDPOINT OF A SOCIAL BEHAVIORIST 154 (1934).
[64] Jack Martin & Bryan Sokol, *Generalized others and imaginary audiences: A neo-Meadian approach to adolescent egocentrism*, 29 NEW IDEAS IN PSYCHOLOGY 364, 369 (2011).
[65] Bernard N. Meltzer, *Mead's Social Psychology*, *in* SYMBOLIC INTERACTION: AN INTRODUCTION TO SOCIAL PSYCHOLOGY 49 (Larry T. Reynolds & Nancy J. Herman eds., 1994).
[66] *Id.*
[67] Mitchell Aboulafia, *George Herbert Mead*, STANFORD ENCYCLOPEDIA OF PHILOSOPHY, available at https://plato.stanford.edu/entries/mead/#RolSelGenOth.
[68] *Id.*

criminal "has not taken on the attitude of the generalized other toward property [and therefore] lacks a completely developed self."[69]

The idea that the "generalized other" perspective sets expectations for one's own behavior resonates with tort law's use of the "reasonable person" standard, which effectively sets the expectations for all members of a society. The "generalized other" concept is related to psychological research on theory of mind,[70] perspective taking,[71] and metacognition,[72] in that it provides a basis for reasoning about what others do, though it is distinct in that it contemplates a group or intersubjective perspective rather than the perspective of a discrete individual.[73] To the extent that jurors call upon the concept of a "generalized other" when applying the reasonable person standard, I expect they will tend to favor a positive interpretation (as defined above). The "generalized other" perspective is a product of experience more likely to be informed by reflections on observed human behavior than by abstract principles like utility maximization.[74]

III. Empirically Testing Lay Constructions of the Reasonable Person

How do lay decision makers interpret and apply the reasonable person standard in tort cases? Do they favor an empirical standard, an aspirational standard, or an economic standard? I conducted a series of four original experiments to investigate these questions. I will note at the

---

[69] Meltzer, *supra* note 65, at 51. Mead has been criticized for oversimplifying the "generalized other" concept by "assuming, apparently, a single, universal generalized other for the members of each society—rather than having a variety of generalized others (even for the same individuals) at different levels of generality." *Id.*

[70] *See, e.g.*, Cybele Raver & Bonnie J. Leadbeater, *The Problem of the Other in Research on Theory of Mind and Social Development*, 36 Human Development 350 (1993).

[71] Martin & Sokol, *supra* note 64.

[72] Mark Sadoski, *Imagination, cognition, and persona*, 10 Rhetoric Rev. 266 (1992).

[73] Martin & Sokol, *supra* note 64.

[74] *Id.*

outset that, given the myriad factors that play into tort cases, I do not expect that any of the three possibilities (empirical, aspirational, and economic) will explain all of the variation in people's negligence verdicts. However, my aim is to identify which of these standards are most frequently reflected in the decisions people make.

Here, I describe the rationale, methodology, and findings of my four experiments. Part III.A describes the first experiment, Part III.B the second, and so on through Part III.D. Each of the four subparts employs the same structure, first providing a brief overview, then providing detailed descriptions of the method and results.

A. Experiment One: Positive or Economic?

1. Overview

My first experiment investigated whether lay decision makers considering negligence cases are more inclined to define the reasonable person in economic terms or in positive terms. In my experiment, participants acted as jurors and rendered verdicts for four hypothetical negligence cases (presented as written vignettes).[75] Each case involved an injured plaintiff who was suing a corporate defendant for failing to take a precaution that would have prevented the plaintiff's injury.[76] Participants received two critical pieces of information about each case. First,

[75] This is common practice in legal scholarship. *See, e.g.*, Justin Sevier, *Testing Tribe's Triangle: Juries, Hearsay, and Psychological Distance*, 103 GEO. L.J. 879 (2014); Avani Mehta Sood, *Attempted Justice: Misunderstanding and Bias in Psychological Constructions of Critical Attempt*, 71 STAN. L. REV. 593 (2019).

[76] The plaintiffs were always individuals and the defendants were always corporations because I did not want to introduce additional variables (i.e. the individual or corporate status of the plaintiff or defendant) into the experiment. Previous research demonstrates that whether defendants, in particular, are individuals or corporations can substantially affect decision makers' verdicts. *See, e.g.*, Valerie P. Hans & M. David Ermann, *Responses to Corporate Versus Individual Wrongdoing*, 13 L. & HUM. BEHAV. 151 (1989); Valerie P. Hans, *The Contested Role of the Civil Jury in Business Litigation*, 79 JUDICATURE 242 (1995); Robert J. MacCoun, *Differential Treatment of*

participants were told how many people or companies in the defendants' position would have chosen to take the relevant precaution. This information (the "positive information") would be relevant if participants applied a positive definition of the reasonable person (i.e. an empirical or aspirational standard). Second, participants were told whether the relevant precaution was cost-effective, as defined by the Hand formula. This information (the "economic information") would be relevant if participants applied an economic reasonable person standard. I found that positive information significantly affected participants' verdicts, while economic information did not.

2. Method

a. Participants

Participants completed the experiment through Amazon Mechanical Turk,[77] an online platform frequently used by researchers to recruit participants.[78] The final sample of 99 participants[79] included 58 men and 41 women, ranging in age from 20 years to 68 years with an average age of 37.16 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Corporate Defendants by Juries: An Examination of the 'Deep-Pockets' Hypothesis, 30 L. & Soc'y Rev. 121 (1996).

[77] Specifically, the study was constructed using IBM's Qualtrics survey software, https://www.qualtrics.com/. Participants were recruited through Amazon Mechanical Turk, where they followed a link to the Qualtrics survey.

[78] *See, e.g.*, Sevier, *supra* note 75; Mehta Sood, *supra* note 75.

[79] Initially, one hundred participants were solicited, but one participant completed the study twice. The participant's initial response was included in the data set and the second response was excluded.

b. Procedure

i. Instructions

After completing a consent form, participants read the following instructions:

> **Your job in this study is to play the role of a juror deciding lawsuits.**
>
> You will read brief scenarios describing the facts of the lawsuits. Assume that all of the facts presented to you are 100% accurate. After reading about each lawsuit, you will be asked to render a decision in the case.
>
> You will make decisions about four different lawsuits. Afterward, you will be asked some additional questions about yourself and your experiences.

Participants were then informed that all of the four lawsuits they would decide were negligence cases, in which an injured person (the plaintiff) claimed that another person or company (the defendant) acted negligently, causing the plaintiff's injury. Participants were instructed to focus only on the plaintiff's claim against that particular defendant, without consideration as to whether the plaintiff may be able to sue someone else.

Finally, participants were presented with representative jury instructions defining negligence (taken from the pattern civil jury instructions for the State of Delaware)[80]:

> Negligence is the lack of ordinary care; that is, the absence of the kind of care a reasonably prudent and careful person would exercise in similar circumstances. That standard is your guide. If a person's conduct in a given circumstance doesn't measure up to the conduct of an ordinarily prudent and careful person, then that person was negligent. On the other hand, if the person's conduct does measure up to the conduct of a reasonably prudent and careful person, the person wasn't negligent.
>
> The mere fact that an accident occurred isn't enough to establish negligence.

[80] Del. P.J.I. Civ. § 5.1. This pattern jury instruction was selected as representative based on a review of the pattern jury instructions presented in Kelly & Wendt, *supra* note 10.

To facilitate comprehension, participants were prompted to answer questions about the experiment's instructions and the legal definition of negligence. Specifically, after reading each set of instructions, participants were asked which of several sentences did not appear in the preceding instructions. In each instance, the choices provided to participants included three key sentences from the instructions (prompting them to re-read those key sentences) and one distractor. Participants had to answer the question correctly to proceed to the next page of the study. Note, however, that participants were not expected to memorize the definition of negligence; following the instruction pages, the definition was reprinted for participants on each page of the experiment.

ii. The Negligence Cases

After reviewing the instructions, participants decided four hypothetical negligence cases, each of which was presented on its own separate page. Each case's page contained (i) a written vignette (322 to 379 words in length) describing the relevant case facts, (ii) pattern jury instructions for the plaintiff's negligence claim, and (iii) four questions for the participant to answer about the case.

Each case vignette followed the same basic pattern.[81] An initial orienting sentence told participants who was suing whom and for what. For example, one opening sentence read as follows: "The plaintiff, Patrick Pendleton, is suing the defendant, Dolman Transportation, claiming that Dolman Transportation's negligence caused him injury." The next ten to twelve sentences described a factual scenario in which the defendant corporation opted not to take a precaution that

[81] Two of the four case vignettes—*Sanders v. A & G Cosmetics* and *Windsor v. International Computers*—were adapted from vignettes used in Daniel Kahneman, David A. Schkade, & Cass R. Sunstein, *Shared Outrage and Erratic Awards: The Psychology of Punitive Damages*, 16 J. RISK & UNCERTAINTY 49 (1998).

would have prevented an injury suffered by the plaintiff. In the example case of *Pendleton v. Dolman Transportation*, the plaintiff suffered burns in a car accident that would have been prevented if Dolman Transportation had purchased and used chemical-hauling trucks with specially-reinforced sides.

The closing sentences of each vignette provided participants with the critical positive information and economic information. For the positive information, participants were directly told what percentage of people or companies would have taken the relevant precaution. For example, in *Pendleton v. Dolman Transportation*, the positive information read as follows: "Given all of the information available at the time, 90% of companies in Dolman Transportation's position would have chosen to buy the brand new, top-of-the-line chemical-hauling trucks with specially-reinforced sides." For the economic information, participants were directly told whether the relevant precautions were cost-effective pursuant to the Hand formula. For example, in *Pendleton v. Dolman Transportation*, the economic information read as follows:

> Dolman Transportation's choice to purchase the older chemical-hauling trucks without side reinforcements led to a 1% greater risk of a side-crash-related explosion. Such an explosion would be expected to cause $5,000,000 of damage.
>
> This means that, if it had purchased the brand new, top-of-the-line chemical-hauling trucks, Dolman Transportation would have been expected to save $50,000 of costs to other members of society. Purchasing the brand new, top-of-the-line chemical-hauling trucks would have cost Dolman Transportation $75,000 more than purchasing the older chemical-hauling trucks.

The order in which the positive information and economic information was provided to participants was counterbalanced: each participant saw the positive information first for two cases

and the economic information first in the other two cases.[82] All four case vignettes can be viewed in Appendix A.

The positive information and the economic information served as the *independent variables* in my experiment. For each case, the positive information was manipulated such that either 10% or 90% of companies in the defendants' position would have taken the relevant precautions, and the economic information was manipulated such that the precaution was either cost-justified (B < PL[83]) or not cost-justified (B>PL[84]). Each possible combination of positive information and economic information was randomly assigned to one (and only one) of the four case vignettes.[85] Thus, each participant decided one case in which both the positive information and the economic information indicated the defendant was negligent (90% of companies would have taken the precaution, which was cost-justified (B<PL)), one case in which both the positive information and the economic information indicated the defendant was not negligent (only 10% of companies would have taken the precaution, which was not cost-justified (B>PL)), one case in which the positive information indicated that the defendant was negligent but the economic information indicated that the defendant was not negligent (90% of companies would have taken the precaution, though it was not cost-justified (B>PL)), and one case in which the positive information indicated that the defendant was not negligent but the economic information indicated that the defendant was negligent (only 10% of companies would have taken the precaution, though it was cost-justified ((B<PL)). This design is summarized in Table 1.1.

[82] The sequence in which the positive information and the economic information were presented had no effect on participants' verdicts, so I pool together positive-then-economic presentations and economic-then-positive presentations in all of the analyses reported herein.
[83] Specifically, in the cost-justified condition, B = .5(PL).
[84] Specifically, in the non-cost-justified condition, B = 1.5(PL).
[85] Order was randomized using IBM's Qualtrics, https://www.qualtrics.com/.

	Economic information indicates negligence	Economic information indicates no negligence
Positive information indicates negligence	90% would have taken precaution Precaution was cost-justified (B<PL)	90% would have taken precaution Precaution was not cost-justified (B>PL)
Positive information indicates no negligence	10% would have taken precaution Precaution was cost-justified (B<PL)	10% would have taken precaution Precaution was not cost-justified (B>PL)

Table 1.1. Summary of experimental design in Experiments One and Two. Each participant acted as a juror for four cases. One case corresponded to each cell in the 2x2 grid above.

After each case vignette, participants reviewed a representative set of negligence jury instructions (the Delaware pattern jury instructions excerpted above).[86] Participants were then prompted to answer four questions about the case.

The first two questions were critical to my analyses. First, participants were asked to render a verdict. Specifically, participants were asked: "Do you find that the defendant, [Defendant's Name], was negligent?" Participants could select one of two options: "No. The defendant was not negligent," or "Yes. The defendant was negligent." Second, participants were asked to rate their confidence in their verdict on a scale from 0 (not at all confident) to 10 (extremely confident). I combined participants' confidence scores with their verdicts to create a continuous measure on a 21-point scale, on which a score of 21 reflected that the participant was extremely confident that the defendant *was* negligent and a score of 0 reflected that the participant was extremely confident

[86] Del. P.J.I. Civ. § 5.1.

that the defendant *was not* negligent. I refer to this 21-point scale as a measure of "perceived negligence" or as a "negligence rating." It served as my primary *dependent variable* (the outcome compared across conditions in my experiment).

Participants were asked two additional questions: whether "most people or companies" in the defendant's position would have taken the relevant precaution (e.g. "Would most people or companies in Dolman Transportation's position have purchased the top-of-the-line trucks with specially-reinforced sides?"), and whether the "reasonably prudent and careful person or company" in the defendant's position would have taken the relevant precaution (e.g. "Would the reasonably prudent and careful person or company in Dolman Transportation's position have purchased the top-of-the-line trucks with specially-reinforced sides?"). I asked these questions to probe the extent to which participants' answers agreed with their verdicts (and whether any dissociations might imply application of an economic or aspirational reasonable person standard).

As noted above, the four questions asked about each case were included on the same page as the case vignette and jury instructions, allowing participants to freely refer back to the vignette and instructions while making their decisions. I opted for this design because my interest is in how participants apply the reasonable person standard to the factual scenarios, and not in how well jurors remember the standard or the scenarios. However, participants could not return to a prior page after completing it. The four cases were presented to each participant in a random order.[87]

iii. Demographic and Wrap-Up Questions

After rendering verdicts for all four cases, participants provided some basic demographic

[87] Order was randomized using IBM's Qualtrics, https://www.qualtrics.com/.

information: gender, their level of education, the highest level of education completed by either of their parents, and their native language. Participants were also asked what they thought the study was testing. The vast majority of participants responded either by saying that they were uncertain, or with a general statement about negligence (e.g. "people's perceptions of negligence" or "how people think about the legal concept of 'negligence'"). Only six of the 99 participants in Experiment One alluded to the manipulations of positive and economic information in their responses. Re-running my statistical analyses with those six participants excluded does not materially alter my results.

c. Hypotheses

The experiment tests whether lay decision makers are more inclined to apply a positive (i.e. empirical or aspirational) or economic definition of the reasonable person when deciding negligence cases. If participants apply a positive definition, then the positive information should substantially affect their negligence ratings. Specifically, participants should rate the defendant as more negligent when 90% of people or companies would have taken the relevant precaution than when only 10% of people or companies would have taken the relevant precaution. If participants apply an economic definition, then the economic information should substantially affect their negligence ratings. Specifically, participants should rate the defendant as more negligent when the precaution was cost-justified (B<PL) than when the precaution was not cost-justified (B>PL). Both of these things could be true, of course. However, I hypothesized that participants would be more likely to define the reasonable person positively than in economic terms, and therefore that participants' negligence ratings would be more influenced by positive information than economic information. I based this hypothesis on two reasons. First, it has been observed that a positive

definition of the reasonable person is likely more accessible to lay decision makers than a utilitarian definition.[88] Second, to the extent the reasonable person standard invites decisionmakers to call upon the concept of a "generalized other," that concept arises from, and is likely defined by, experience rather than abstract logical principles.[89]

3. Analyses and Results

The critical inquiry was whether and how positive information and economic information influenced participants' negligence ratings (measured by the 21-point scale described above). The analyses reported below demonstrate (i) that positive information significantly influenced participants' negligence ratings, (ii) that economic information did not significantly influence participants' negligence ratings, and (iii) that the influence of positive information on participants' negligence ratings was significantly greater than the influence of economic information on participants' negligence ratings.[90] This conclusion is further supported by a follow-up analysis examining how positive information and economic information affected participants' ratings in the first scenario to which each participant responded.

In addition to these focal analyses, I briefly describe how participants' negligence verdicts correspond to their answers to the other two questions asked of them (i.e. what "most people or companies" would have done and what the "reasonably prudent and careful person or company" would have done under the circumstances). I also describe how participants' negligence ratings varied across cases.

---

[88] *See* Heidi Li Feldman, *Prudence, Benevolence, and Negligence: Virtue Ethics and Tort Law*, 74 CHI.-KENT L. REV. 1431, 1433 (1998).
[89] *See* Martin & Sokol, *supra* note 64.
[90] For a contingency table summarizing participants' binary verdicts in each condition, see Appendix A.

a. Positive Information, but not Economic Information, Influences Negligence Ratings

A two-way repeated-measure ANOVA revealed that positive information significantly affected participants' negligence ratings, $F(1, 98) = 38.184$, $p < .001$. As shown in Figure 1.1, participants deciding cases in which 90% of people or companies in the defendants' position would have taken the relevant precaution perceived the defendant as significantly more negligent ($M = 15.242$, $SD = 5.742$) than participants deciding cases in which only 10% of people or companies would have taken the relevant precaution ($M = 10.268$, $SD = 6.034$), $t(98) = 6.179$, $p < .001$, $d = .621$. Economic information did not significantly affect participants' perceptions of negligence,[91] nor did any interaction between positive information and economic information.[92] Positive information affected participants' perceptions of negligence significantly more than economic information did, $t(98)=3.619$, $p < .001$, $d = .364$.[93]

[91] $F(1,98)=2.279$, $p=.134$. Participants' mean negligence rating when the defendant was cost-justified was 13.308 ($SD = 5.596$); participants' mean negligence rating when the defendant was not cost-justified was 12.202 ($SD = 5.707$).

[92] $F(1,98)=3.074$, $p=.083$.

[93] To test the relative influence of positive information and economic information, I calculated and compared differences across levels of each (positive information difference = mean negligence rating in 90% cases – mean negligence rating in 10% cases; compared with economic information difference = mean negligence rating for cost-justified cases – mean negligence rating for non-cost-justified cases).

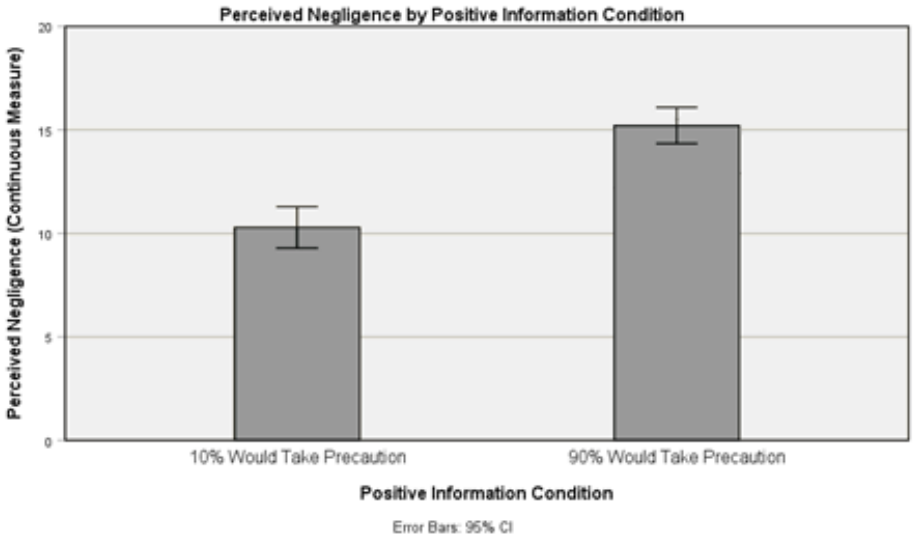


Figure 1.1. Mean negligence rating by positive information condition in Experiment One.

An alternative method for analyzing this data uses participants' binary verdicts, categorizing participants by pattern. Each participant judged four cases, so there were 16 potential patterns of verdict across cases (2^4 = 16). I was particularly interested in two patterns. First, participants who always found the defendant negligent in the 90% positive information condition, and never in the 10% positive information condition, gave responses consistent with a pure positive standard. Second, participants who always found the defendant negligent when the relevant precaution was cost-justified, and never when the relevant precaution was not cost-justified, gave responses consistent with a pure economic standard. Among the subset of participants whose verdicts matched one of these two patterns, participants were significantly more likely to respond consistent with a pure positive standard (N = 19) than with a pure economic standard (N = 5), χ^2(1 df)=8.167, p=.004.

While this alternative analysis also supports the positive hypothesis, it must be noted that only 24 of 99 participants gave responses consistent with deciding the cases on positive

information or economic information alone. The other 75 participants were being influenced by some combination of those factors and other factors. One noteworthy additional factor in this study was a general pro-plaintiff bias: participants were significantly more likely than chance to find for the plaintiff in all four cases (N = 16), $\chi^2(1\ df)=16.599$, $p < .001$.[94] Given that the cases all involve corporate defendants, I suspect this pro-plaintiff bias may be a manifestation of a previously-documented bias against corporate defendants.[95]

Table 1.2 shows summarizes patterns in participants' negligence verdicts. Fifty-five of 99 participants' responses fell into my catch-all "other" category—their responses were not consistent with a pure positive standard, a pure economic standard, or ruling exclusively for the plaintiff or for the defendant. The fact that a majority of participants fall into this catch-all category is a useful reminder that myriad considerations bear on participants' negligence judgments.[96]

Verdict Pattern	# of Participants
Pure Positive Standard	19
Pure Economic Standard	5
Always Rule for Plaintiff	16
Always Rule for Defendant	4
Other	55

Table 1.2. Participants' verdict patterns in Experiment One.

[94] I observed no significant pro-defendant bias, $\chi^2(1\ df)=0.825$, $p=.363$.
[95] *See* note 76, *supra*.
[96] *See* Votruba, *supra* note 20.

b. Follow-Up Analysis with First Case Only

A skeptic could conceivably argue that the significant relationship I observe between positive information condition and negligence ratings is an artifact of the repeated-measure design. Each participant responded to four case vignettes. The positive information manipulation could have captured participants' attention. Thus, participants might have rated defendants as more negligent in the 90% condition than the 10% condition due to a comparison of information across cases. To address this possibility, I compiled and analyzed a data set that included only participants' verdicts for the *first case they encountered.* Even when analyzing only the initial verdicts, participants rated the defendant as significantly more negligent in the 90% condition (M = 15.209, SD = 7.507) than the 10% condition (M = 9.214, SD = 8.416), $t(97) = 3.680$, $p < .001$.[97] This indicates that the influence of positive information on negligence verdicts was present from the outset of the study. (Economic information did not affect negligence ratings in this restricted data set, $t(97) = 1.240$, $p = .218$.)

c. Responses to the Most People and Reasonable Person Questions

After providing negligence verdicts and confidence ratings in each case, participants were asked whether "most people or companies" would have taken the relevant precaution, and whether the "reasonably prudent and careful person or company" would have done what the defendant failed to do. These questions provided an opportunity to check whether people's answers corresponded to their verdicts. If participants apply an empirical definition of the reasonable person, one would expect the three responses to agree: people's verdicts should reflect their answer

[97] The same pattern holds true if participants' verdicts are coded dichotomously, $\chi^2(1 \text{ df}, N = 99) = 10.96$, $p = .001$.

to the "reasonable person" question, which under an empirical reasonable person standard should also match their answer to the "most people or companies" question. On the other hand, if participants' responses to the "most people or companies" question consistently diverged from the other responses, it could suggest that participants are using an economic or aspirational definition.

In the vast majority of cases (74.49%), participants' verdict, their answer to the reasonable person question, and their answer to the "most people" question were all in agreement. In only 9.8% of cases did participants' negligence verdict agree with their response to the reasonable person question but diverge from their response to the most people question (as one might expect if participants were applying an economic or aspirational interpretation of the reasonable person). These data are at least consistent with participants applying an empirical definition of the reasonable person standard—a notion I explore further in Experiments Three and Four.

d. The Effect of Scenario on Negligence Ratings

Finally, it is worth noting that participants' negligence ratings varied significantly across the four case vignettes, $F(3, 98) = 9.500$, $p < .001$. Descriptively, participants' mean negligence ratings were as follows: *Sanders v. A & G Cosmetics*, M = 14.889 (SD = 7.065); *Windsor v. International Computers*, M = 14.263 (SD = 7.604); *Vaughan v. Menlove Farms, Inc.*, M = 12.030 (SD = 8.151); and *Pendleton v. Dolman Transportation*, M = 9.838 (SD = 8.104). This variability is unsurprising, given the wide variety of facts in each case that could make participants more or less likely to find the defendant negligent. As noted above, only 24 participants gave responses that perfectly matched either the positive reasonable person standard or the economic reasonable person standard, indicating that most participants were influenced by a combination of these factors and other factors. It is important to note, however, that participants' ratings were not at

floor or ceiling for any of the four cases, meaning there was room for positive and economic information to influence participants' negligence ratings in all cases.

B. Experiment Two: Positive or Economic? (Holding Manipulation Magnitude Constant)

1. Overview

In Experiment One, positive information influenced participants' perceptions of the defendant's negligence, while economic information did not. However, Experiment One had one important limitation: the *magnitudes* of the manipulations of positive and economic information were not consistent. With respect to positive information, either 90% or 10% of entities in the defendants' position would have acted differently—a 9:1 difference between conditions. With respect to economic information, however, the cost of the precaution was either 50% or 150% of the expected cost of the failure to take precautions—effectively a 3:1 difference between conditions. If participants were applying the economic reasonable person standard, in strictest terms, then this manipulation of economic information should have been sufficient to swing the case. However, it is plausible that the difference in magnitudes caused participants to place more weight on the positive information, relative to the economic information, than they otherwise would have.

For these reasons, Experiment Two was designed to replicate Experiment One while holding the magnitude of manipulations constant. For the positive information manipulation, I again used 90% and 10%, but for the economic information manipulation, I used 9:1 ratios. For example, if the cost of the precaution in a vignette equals $100,000, then the expected cost of failing to take the precaution would be $900,000 in the cost-justified condition (B<PL), or $11,111 in the non-cost-justified condition (B>PL).

Aside from this change, there was one other minor difference between experiments. In Experiment Two, I added four attention check questions, spread throughout the study. Each question was a basic, four-alternative, forced-choice question about what they had read on the previous screen. I set an *a priori* exclusion criterion that participants who missed more than two of the four attention checks would be excluded.

Aside from these differences, the method used in Experiment Two was identical to that in Experiment One.

2. Method

a. Participants

111 participants were recruited using Amazon Mechanical Turk.[98] Thirteen of these participants were excluded based my *a priori* attention check criterion, leaving 98 for analysis. The final sample of 98 included 55 men, 42 women, and one participant who preferred not to identify. Participants ranged in age from 20 years to 69 years with an average age of 35.64 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

b. Procedure

The procedure used in Experiment Two was identical to that used in Experiment One, subject to the variations described in the experiment overview above (Part III.B.1).

[98] Initially, 120 participants were solicited, but nine responses were excluded because they came, or could have come, from the same participant as an earlier response.

c. Hypotheses

Experiment Two tested the same hypotheses as Experiment One.

3. Analyses and Results

My analyses in Experiment Two largely paralleled those in Experiment One. As in Experiment One, my examination of the influences of positive and economic information on participants' negligence ratings revealed (i) that positive information significantly influenced perceived negligence, (ii) that economic information did not, and (iii) that the influence of positive information on perceived negligence was greater than any influence of economic information. [99]

a. Positive Information, but not Economic Information, Influences Negligence Ratings

As in Experiment One, a two-way repeated-measure ANOVA revealed that positive information significantly affected participants' negligence ratings, $F(1, 97) = 18.339$, $p < .001$. Participants rated defendants as significantly more negligent in cases in the 90% condition ($M = 15.413$, $SD = 5.116$) than in the 10% condition ($M = 11.862$, $SD = 6.130$), $t(97) = 4.282$, $p < .001$, $d = .433$, as shown in Figure 1.2. Economic information had no effect,[100] nor did the interaction of positive and economic information.[101] Positive information affected participants' perceptions

[99] For a contingency table summarizing participants' binary verdicts in each condition, see Appendix A.

[100] $F(1,98) = .016$, $p = .899$. Participants' mean negligence rating when the defendant was cost-justified was 13.592 ($SD = 5.304$); participants' mean negligence rating when the defendant was not cost-justified was 13.684 ($SD = 5.220$).

[101] $F(1,98) = .851$, $p = .359$

of negligence significantly more than economic information did, $t(97)= 3.542$, $p < .001$, $d = .358$.[102]

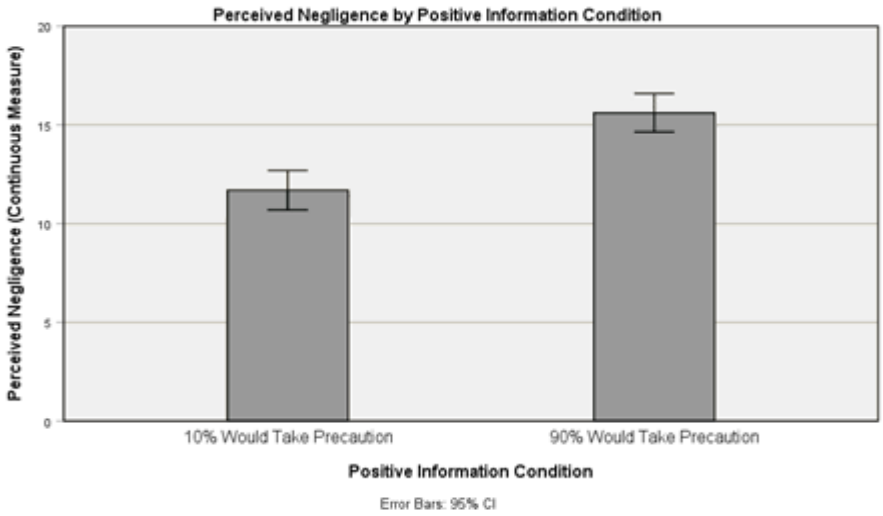


Figure 1.2. Mean negligence rating by positive information condition in Experiment Two.

I conducted an alternative analysis categorizing individual participants based on the pattern of their verdicts, using the same categories as Experiment One. Participants' verdict patterns are summarized in Table 1.3.

[102] For explanation of this contrast, see note 92, *supra*.

Verdict Pattern	# of Participants
Pure Positive Standard	15
Pure Economic Standard	4
Always Rule for Plaintiff	21
Always Rule for Defendant	0
Other	58

Table 1.3. Participants' verdict patterns in Experiment Two.

Among the subset of participants who responded consistent with either a pure positive standard or a pure economic standard, participants were significantly more likely to respond consistent with the positive standard, $\chi^2(1 \text{ df}) = 6.368$, p=.012. However, as in Experiment One, the majority of participants' verdict patterns were not consistent with either a pure positive standard or a pure economic standard. Finally, it merits mention that, similar to Experiment One, participants were more likely than chance would predict to simply rule for the plaintiff in each case, $\chi^2(1 \text{ df}) = 38.533$, $p < .001$.

b. Follow-Up Analysis with First Case Only

I ruled out the possibility that my findings are an artifact of my within-participants design by compiling and analyzing a data set comprising only participants' first responses. Even when analyzing only the initial verdicts, participants rated the defendant as significantly more negligent in the 90% condition (M = 17.174, SD = 5.555) than the 10% condition (M = 12.404, SD = 7.959), $t(96) = 3.397$, $p < .001$, $d = .688$.[103] This indicates that the influence of positive information on

[103] The same pattern holds true if participants' verdicts are coded dichotomously, $\chi^2(1 \text{ df}, N = 99)=10.96$, p=.001.

negligence verdicts was present from the outset of the study. (Economic information did not affect negligence ratings in this restricted data set, $t(96) = 1.654$, $p = .101$.)

c. Responses to the Most People and Reasonable Person Questions

Similar to Experiment One, participants' verdict agreed with their answers to the reasonable person question and the "most people" question in the vast majority of cases (72.2%). Only in 9.9% of cases did participants' negligence verdict agree with their response to the reasonable person question, but disagree with their response to the "most people" question (the pattern of responses that would most strongly suggest application of an economic or aspirational reasonable person standard).

d. The Effect of Scenario on Negligence Ratings

Participants' negligence ratings varied significantly across the four case vignettes, $F(3, 97) = 8.537$, $p < .001$, and were strikingly similar to those in Experiment One: *Sanders v. A & G Cosmetics*, M = 15.969 (SD = 6.730); *Windsor v. International Computers*, M = 15.020 (SD = 7.346); *Vaughan v. Menlove Farms, Inc.*, M = 11.918 (SD = 7.954); and *Pendleton v. Dolman Transportation*, M = 11.633 (SD = 8.065). The similarities to Experiment One underscore the idea that while positive information is influencing participants' negligence judgments, there are other factors systematically affecting those judgments too. The vignettes did not change from Experiment One to Experiment Two. The results indicate that whatever latent factors in the vignettes are contributing to negligence ratings had largely consistent effects across experiments.

C. Experiment Three: Empirical or Aspirational?

1. Overview

Experiments One and Two provide evidence that lay decision makers are influenced by positive information (about the proportion of people or companies that would have taken precautions), and the influence of this information is greater than any influence of economic information (about whether precautions were cost-justified). This suggests that lay decision makers' operationalization of the reasonable person standard is more of a positive standard than an economic one. This finding leads to an immediate follow-up question: which positive variant of the reasonable person standard do lay decision makers favor, the empirical definition or the aspirational definition? Experiments Three and Four explored this question.

In Experiment Three, participants were again asked to act as jurors, reviewing written vignettes of negligence cases and rendering verdicts. As in Experiments One and Two, each case involved an injured plaintiff suing a corporate defendant for failing to take specific precautions.[104] Further, like the prior experiments, participants were given relevant positive information—information specifying what percentage of people or companies would have taken the precautions the defendant did not take. The critical feature of Experiment Three was that the positive information could reflect any of *five* different conditions, in which anywhere from 0% to 90% of companies would take the relevant precaution under the circumstances.[105] Manipulating this percentage across five levels I allowed me to investigate where positive information begins to affect negligence verdicts along this continuum.

[104] The case vignettes used in Experiment Three were, for the most part, identical to those used in Experiments One and Two, with the exceptions described in the Method section below.

[105] In Experiment Three, the case vignettes did not include any economic information about the cost-effectiveness of precautions.

2. Method

a. Participants

Sixty (60) participants completed Experiment Three using Amazon Mechanical Turk.[106] Participants included 28 men, 31 women, and one participant who preferred not to identify.[107] Participants ranged in age from 20 years to 71 years, with an average of 39.2 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

b. Procedure

Experiment Three's procedure generally resembled the prior experiments, with three exceptions. First, in Experiment Three, participants responded to five negligence cases rather than four. The five cases included the four cases used in the prior studies, plus one additional case, *Lawson v. TGI International*, adapted from previous research.[108]

Second, Experiment Three tested one independent variable (positive information), rather than two (positive information and economic information). Thus, the vignettes used in Experiment Three did not include any economic information about the cost-effectiveness of the precautions the defendant declined to take. Rather, every vignette ended with positive information directly informing participants what percentage of people or companies would have chosen to take the

[106] Specifically, the study was constructed using IBM's Qualtrics survey software, https://www.qualtrics.com/, and participants were recruited through Amazon Mechanical Turk, where they followed a link to the Qualtrics survey.

[107] No participants were excluded from the sample in Experiment Three.

[108] The *Lawson v. TGI International* vignette was adapted from Daniel Kahneman, David A. Schkade, & Cass R. Sunstein, *Shared Outrage and Erratic Awards: The Psychology of Punitive Damages*, 16 J. RISK & UNCERTAINTY 49 (1998). The vignette can be viewed in Appendix A.

precautions (I refer to this percentage as the "population performance percentage" or "PPP"). The vignettes used in Experiment Two ranged from 190 to 281 words in length.

Third, in Experiment Three, the manipulation of positive information includes five conditions, corresponding to five possible levels of PPP: 0%, 10%, 25%, 50%, and 90%. I chose the levels based on the implications of the aspirational and empirical hypotheses. Specifically, values are concentrated on the low end of the continuum (0%, 10%, and 25%) to allow for evaluation of the aspirational hypothesis: that hypothesis suggests that participants should not find defendants negligent in the 0% condition, but should begin find them negligent at some low PPP threshold (perhaps around 10%), after which the effect of PPP should taper. The empirical hypothesis, in contrast, suggests that the biggest effects of PPP should arise at some point around 50% PPP (perhaps right at 50%, perhaps 50.0001%, perhaps 51%). I did not include a 75% condition because I am unfamiliar with any theoretical basis for predicting that the difference between 51% PPP and 75% PPP should affect participants' negligence judgments. (Such a difference certainly might affect negligence verdicts, but it would not be predicted by either an empirical or aspirational reasonable person standard as I have defined them).

Each level of PPP was randomly assigned to one (and only one) of the five case vignettes.[109] Thus, each participant decided one case with a PPP of 0%, one case with a PPP of 10%, and so on, but which case featured which level of PPP varied across participants.

I adapted the instructions participants reviewed in Experiment Three to the extent needed to reflect these procedural changes. Otherwise, the instructions used in Experiment Three were identical to those used in Experiments One and Two.

Experiment Three employed the same dependent measures as prior studies.

[109] Specifically, order was randomized using IBM's Qualtrics, https://www.qualtrics.com/.

c. Hypotheses

The experiment was designed to probe whether participants are more inclined to use an empirical or aspirational understanding of the reasonable person standard in evaluating negligence cases.

If people apply an empirical reasonable person standard, then, by definition, whether the defendant was negligent depends on what most people would have done under the circumstances. On this view, PPP should influence on participants' negligence ratings, with participants perceiving the defendant as more negligent when the PPP is high (when the majority of people, e.g. 90%, would have taken the relevant precautions) than when the PPP is low (when only a minority of people, e.g. 25%, would have taken the precautions).

If, on the other hand, people apply an aspirational reasonable person, then the reasonable person is expected to do what the very best among us would do in a situation. On this view, the effects of PPP on negligence ratings should be concentrated at relatively low levels of PPP. So long as a (potentially-quite-small) minority of people—for the purposes of this Chapter, I have assumed about 10%—would have taken the relevant precautions, the defendant is negligent. Once PPP exceeds the aspirational threshold, further changes in PPP should not greatly influence participants' perceptions of negligence.

In sum, whether people generally apply an empirical or aspirational reasonable person standard, one would predict a substantial jump in negligence ratings at some point along the PPP continuum. But the two possibilities yield different predictions about *where* that substantial jump will occur. If decision makers favor the empirical reasonable person standard, one would expect the proportion of participants who find the defendant negligent to remain flat or perhaps increase slowly with PPP until some point around 50% PPP, where it will jump (as it becomes clear that

most people would have taken the relevant precautions). If decision makers favor the aspirational reasonable person standard, however, the jump should occur much earlier, at some point between 0% PPP and around 10% PPP, as it becomes clear that the best among us would have succeeded where the defendant failed.

Thus, the critical comparisons in Experiment Two are (i) between participants' negligence determinations at 25% and 90% PPP, and (ii) among participants' negligence determinations at 0%, 10%, and 25% PPP. If comparison (i) reveals a significant difference whereas comparison (ii) does not, then Experiment Two will provide evidence for the empirical reasonable person standard over the aspirational reasonable person standard. Specifically, this pattern of results would be incompatible with the aspirational reasonable person standard because it would indicate that PPP differences between 0%, 10%, and 25% are not especially meaningful for participants' negligence ratings, whereas changes at higher PPP thresholds are more influential.

3. Analyses and Results

The analyses reported below demonstrate that PPP significantly affects participants' negligence ratings. The contours of the effect indicate that participants apply something closer to an empirical reasonable person standard than an aspirational reasonable person standard, though the evidence is less clear than in previous experiments.

My primary analysis reveals that participants' negligence ratings increase significantly from PPP=10% to PPP=90%, but do not vary between PPP=0% and PPP=25%; these findings are consistent with the empirical reasonable person hypothesis, but not with the aspirational

hypothesis.[110] And, as in prior studies, participants' assessments of what "most people" would do and what the "ordinary, reasonable person" would do are both closely related to participants' negligence verdicts, consistent with the empirical reasonable person hypothesis. However, my alternative analysis examining patterns in participants' binary negligence verdicts was inconclusive in this experiment.

a. The Effect of PPP on Negligence Ratings

A one-way repeated-measure ANOVA revealed that PPP had a significant effect on negligence ratings, $F(4, 236) = 6.651$, $p < .001$. *Post hoc* comparisons among the different PPP levels revealed that participants in the 90% PPP condition perceived defendants as significantly more negligent ($M = 15.767$, $SD = 5.806$) than participants in each of the 0% ($M = 10.850$, $SD = 6.854$), 10% ($M = 11.800$, $SD = 6.556$), and 25% ($M = 12.033$, $SD = 6.628$) conditions.[111] There were no significant differences among the 0%, 10%, and 25% conditions. Thus, the critical comparisons identified in Part III.C.2.c above yield evidence consistent with the empirical reasonable person hypothesis (negligence ratings at 90% PPP were significantly higher than at 25% PPP), and inconsistent with the aspirational reasonable person hypothesis (no differences between 0% PPP and 25% PPP). Figure 1.3 shows participants' perceptions of the defendant's negligence ratings for each of the five PPP conditions.

[110] For a contingency table summarizing participants' binary verdicts in each condition, see Appendix A

[111] 0% PPP condition: $t(59) = 4.223$, Bonferroni-corrected $p < .001$, $d = .545$; 10% PPP condition: ($t(59) = 3.633$, Bonferroni-corrected $p = .006$, $d = .469$); 25% PPP condition: $t(59) = 2.951$, Bonferroni-corrected $p = .045$, $d = .381$. In addition, *post hoc* tests identified a difference between the 50% PPP condition ($M = 14.150$, $SD = 6.197$) and the 0% condition, $t(59) = 3.233$, Bonferroni-corrected $p = .020$, $d = .417$.

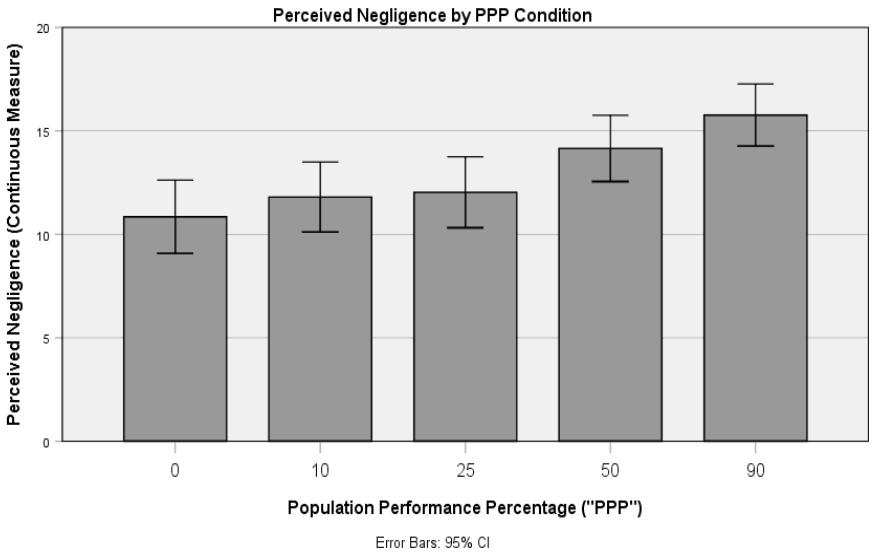


Figure 1.3. Mean negligence rating by PPP condition in Experiment Three.

While I interpret my analysis as providing some support for the empirical reasonable person hypothesis, some caution is required. First, because the PPP levels were not evenly spaced between 0 and 90 (for reasons described above), it appears possible that the relation between negligence ratings and PPP may be linear across the PPP continuum, rather than exhibiting a "jump" around 50%. Second, it bears emphasis that in any event, participants are certainly not applying a "pure" empirical reasonable person standard. Participants' negligence rating is above the midpoint of the scale even in the 0% PPP condition—when no company in the defendant's position would have taken the precaution. As suggested above, I suspect much of this may be due to the fact that, in all cases, the defendants are corporations.[112] I will revisit this issue in Experiment Four.

While my initial analysis provided some support for the empirical standard, my alternative analysis of verdict patterns did not. As in prior studies, I categorized individual participants based

[112] *See* note 76, *supra*.

on the pattern of their verdicts. I was particularly interested in identifying the point (if any) on the PPP spectrum where participants "tipped" from finding the defendant not negligent to finding the defendant negligent. Thus, I identified participants whose verdicts were consistent with the idea of a purely positive standard with a PPP tipping point—participants who found the defendant was not negligent at PPP levels equal to or below a certain threshold, but found the defendant was negligent at each point above that threshold. I categorized participants' responses as reflected an empirical standard if their responses were consistent with a PPP threshold above 49.999%.[113] I categorized participants as reflecting an aspirational standard if their responses were consistent with a PPP threshold between 0% and 24.999%.[114] Looking only at these two categories, the distribution of participants across categories—five participants in the empirical category versus 10 in the aspirational category—did not differ significantly from chance, $\chi^2(1\ df) = 1.667$, $p = .197$. Table 1.4 provides a summary of participants' verdict patterns.

[113] This category included two of the 32 possible verdict patterns: (1) participants who found the defendant was not negligent in the 0%, 10%, and 25% PPP conditions, but was negligent in the 50% or 90% PPP conditions; and (2) participants who found the defendant was not negligent in the 0%, 10%, 25%, or 50% PPP conditions, but was negligent in the 90% condition.

[114] This category also included two of the 32 possible verdict patterns: (1) participants who found the defendant was not negligent in the 0% PPP condition but was negligent in all other PPP conditions; and (2) participants who found the defendant was not negligent in the 0% and 10% PPP conditions, but was negligent in all other PPP conditions.

Verdict Pattern	# of Participants
Aspirational: 0%-10% Tipping Point	3
Aspirational: 10%-25% Tipping Point	7
Empirical: 25%-50% Tipping Point	4
Empirical: 50%-90% Tipping Point	1
Always Rule for Plaintiff	8
Always Rule for Defendant	1
Other	36

Table 1.4. Participants' verdict patterns in Experiment Three.

Two other points merit mentioning. First, as in Experiments One and Two, participants were more likely than chance to rule for the plaintiff in all cases, $\chi^2(1\ df)=20.654$, $p<.001$, which might reflect a bias against corporate defendants. Second, as in prior experiments, a majority participants were influenced by considerations beyond PPP alone.

In sum, Experiment Three produces mixed evidence on whether participants tend to apply a more positive or more aspirational standard. Viewed in the aggregate, participants' negligence ratings appear more consistent with the empirical hypothesis than the aspirational hypothesis: I observed no jump at the lower end of the PPP spectrum. Yet, descriptively, more participants' verdict patterns reflected purely aspirational standards than purely empirical standards. One reason for the mixed results may be that some participants are more inclined to apply an aspirational reasonable person standard to corporations than to individual defendants. This could be the case due to a general bias against corporations—perhaps people just want corporations to

pay (at least when they are sued by individual plaintiffs).[115] Alternatively, it could be because all the case vignettes used in Experiment Three involved corporate defendants making deliberative decisions. Participants might think the corporate defendants should be held to a high standard in such cases, as they often have the opportunity to involve multiple decisionmakers and access to expertise. On this view, participants who would generally conceptualize the reasonable person in empirical terms may have "bumped up" their standard due to the particulars of this set of cases. I return to this idea in Experiment Four.

b. Follow-Up Analysis with First Case Only

In Experiments One and Two, I conducted separate analyses of the participants' first responses that ruled out the possibility that the effects of PPP were artifacts of my repeated-measures design. In this experiment, PPP did not have a statistically significant effect on participants' first negligence ratings, $F(4, 55) = 0.566$, $p = .688$. However, the data set afforded limited statistical power, as the 60 relevant responses were spread over five PPP conditions. Descriptively, participants' first negligence ratings were highest in the 90% PPP condition ($M = 16.400$, $SD = 5.147$), second-highest in the 50% PPP condition ($M = 14.500$, $SD = 6.099$), and lower in the other three conditions (all M's < 13.857).

c. Responses to the Most People and Reasonable Person Questions

I also examined the relationship between participants' answers to the "most people" and "reasonable person" questions and their negligence verdicts. Consistent with prior experiments, participants' verdicts agreed with their answers to both the reasonable person question and the

[115] *Id.*

"most people" question in a substantial majority of cases (69%). In 17.3% cases, participants' negligence verdicts were consistent with their answers to the reasonable person question but not the most people question (which might reflect an application of an aspirational standard); in 4% of cases, participants' negligence verdicts were consistent with their answers to the most people question but not the reasonable person question; and in 9.7% cases participants' answers to the reasonable person and most people questions agreed but differed from their negligence verdicts.

d. The Effect of Scenario on Negligence Ratings

In Experiment Three, mean negligence ratings were 12.267 (SD = 6.889) in *Windsor v. International Computers*, 11.850 (SD = 6.814) in *Vaughan v. Menlove Farms, Inc.*, 10.450 (SD = 7.072) in *Pendleton v. Dolman Transportation*, 14.133 (SD = 5.899) in *Sanders v. A & G Cosmetics*, and 15.900 (SD = 5.018) in *Lawson v. TGI International, Inc.*, $F(4, 236) = 7.632$, $p < .001$.

D. Experiment Four: Empirical or Aspirational for Individual Defendants?

1. Overview

The first three experiments presented in this Chapter indicate that lay interpretation of the reasonable person standard is more positive than economic, and further suggest that participants' use of positive information is more consistent with an empirical interpretation than an aspirational interpretation (though evidence on the latter point is less clear). However, the previous experiments explored participants' constructions of the reasonable person in a relatively homogeneous set of hypothetical cases. Every case involved a corporate defendant, and every case hinged on whether the defendants' deliberate decision not to take a precaution was reasonable.

As mentioned in Part III.C.3.a above, there are reasons to think these features might have nudged some participants to apply a more aspirational standard than they otherwise would have.

Negligence cases, of course, can involve individual defendants rather than corporations, and negligence can arise from failures of a number of cognitive processes *other than* decision making.[116] Experiment Four investigates how the reasonable person standard is construed in this broader universe of cases.

The structure of Experiment Four was identical to that of Experiment Three, but five new case vignettes were substituted for those used in Experiment Three. In each of the new case vignettes, an individual defendant experiences a critical cognitive failure (e.g. a driver failing to see a pedestrian), which results in an injury to an individual plaintiff. As in the previous experiments, participants were told precisely what percentage of the population would have succeeded where the defendant failed[117] (e.g. 10% of people would have seen the pedestrian in time to avoid the accident). I found that participants' verdicts in these cases were most consistent with the empirical reasonable person standard, and that this pattern largely held across different types of cognitive failures.

[116] *See* Rachlinski, *supra* note 1, at 1056 ("If the reasonable person, using her attention, memory, and perceptual abilities would have avoided an accident, then the fact that an accident occurred implies that the actor was engaged in unreasonable conduct."); Jaeger, Levin, & Porter, *supra* note 33, at 263 (discussing how negligence cases might arise from visual failures).

[117] I use the words "failure" and "fail" to refer broadly to the defendant's conduct because, in each scenario, if the defendant had perceived, remembered, or chosen differently, the plaintiff's injury would have been avoided. Of course, one could argue that where the defendant does not perceive, remember, or choose something that only a very small percentage of the population would have perceived, remembered, or chosen, this does not constitute a "failure" at all.

2. Method

a. Participants

A sample of 53 participants (39 men and 14 women)[118] completed this experiment through Amazon Mechanical Turk.[119] Participants ranged in age from 21 years to 71 years, with an average age of 33.09 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

b. Procedure

Procedures were identical to those used in Experiment Three except that the case vignettes were replaced with five new case vignettes. The new vignettes were designed to capture a variety of types of cognitive failures that could potentially underlie a negligence claim. Specifically, one case hinged on the defendant's visual attention and perception, one on the defendant's auditory attention and perception, one on the defendant's memory, one on the defendant's information processing and reaction time, and one on the defendant's decision making.[120] Including case vignettes that involved a variety of cognitive failures allowed me to examine whether participants'

[118] Initially, sixty participants were solicited, but several participants completed the study more than once, resulting in the exclusion of seven responses. Specifically, three participants completed the study multiple times. In each case, the participant's initial response was included in the data set, and all of his or her subsequent responses were excluded.

[119] Specifically, the study was constructed using IBM's Qualtrics survey software, https://www.qualtrics.com/. Participants were recruited through Amazon Mechanical Turk, where they followed a link to the Qualtrics survey.

[120] The decision making case was an adaptation of the *Pendleton v. Dolman Transportation* vignette used in the previous two experiments, revised to streamline the scenario and to present the defendant as a sole proprietor named Donald Dolman.

approach to defining the reasonable person was consistent across a variety of cognitive characteristics.[121]

Just like Experiment Three, the independent variable in Experiment Four was PPP. Participants again decided one case in each of five PPP conditions (0%, 10%, 25%, 50%, or 90%), with the pairing of cases and PPP conditions determined randomly.[122] As in Experiment Three, the PPP manipulation appeared in the final sentence of the case vignette. For example, one closing sentence read as follows: "Assume it is a fact that, given the conditions at the time of the accident, 90% of drivers in Dan's position would have seen Paul." Complete copies of all five case vignettes are included in Appendix A.

The five case vignettes were presented to each participant in a random order.[123] Participants responded to the same four questions about each case as in previous experiments: participants provided a negligence verdict, a rating of confidence in that verdict, an assessment of whether most people would have succeeded where the defendant failed, and an assessment of whether the reasonably prudent and careful person would have succeeded where the defendant failed.

c. Hypotheses

This experiment followed up on Experiment Three by investigating lay application of the reasonable person standard to individual defendants who experienced an assortment of cognitive failures. I hypothesized that participants' verdicts would be more indicative of an empirical, rather

[121] If I had used the same cognitive process in all cases (e.g., if all five cases had hinged on the defendant's failure to see something), my conclusions would necessarily be limited accordingly.
[122] Randomization was done through IBM's Qualtrics, https://www.qualtrics.com/.
[123] *Id.*

than an aspirational, conception of the standard. Indeed, if anything, cases involving failures of basic cognitive processes perception, awareness, memory, and the like seem to call out for an empirical reasonable person standard *more* than cases involving conscious, deliberative decisions about precautions. Assuming that the reasonable person standard should be one with which "all can, if they try, conform,"[124] people are much less able to exert control over processes like vision and memory than decision making. Considering that participants in Experiment Three did not flock to an aspirational standard—even in cases involving deliberate decision making by corporate entities[125]— I expected participants in Experiment Four to evaluate negligence cases using more of an empirical standard than an aspirational one.

3. Analyses and Results

My primary and secondary analyses paralleled those used in Experiment Three. My findings indicate that participants in Experiment Four conceptualized the reasonable person standard in more empirical terms than aspirational terms.[126]

a. The Effect of PPP on Negligence Ratings

A repeated-measure ANOVA confirmed that positive information about human behavior (i.e. PPP) had a significant effect on participants' negligence ratings, $F(4, 208) = 9.789$, $p < .001$. *Post hoc* comparisons among the different PPP levels revealed that participants in the 90% PPP condition ($M = 13.736$, $SD = 6.884$) perceived defendants as significantly more negligent than

[124] Rachlinski, *supra* note 1, at 1060-61 (describing how way the reasonable person standard incentivizes behavior).
[125] *See* Part III.B, *supra*.
[126] For a contingency table summarizing participants' binary verdicts in each condition, see Appendix A.

participants in any of the other four conditions (all t's ≥ 2.997, all Bonferroni-corrected p's ≤.042, all d's ≥ .412),[127] and that the other four PPP conditions did not differ significantly from one another. Figure 1.4 shows participants' perceptions of the defendant's negligence in each of the five PPP conditions.

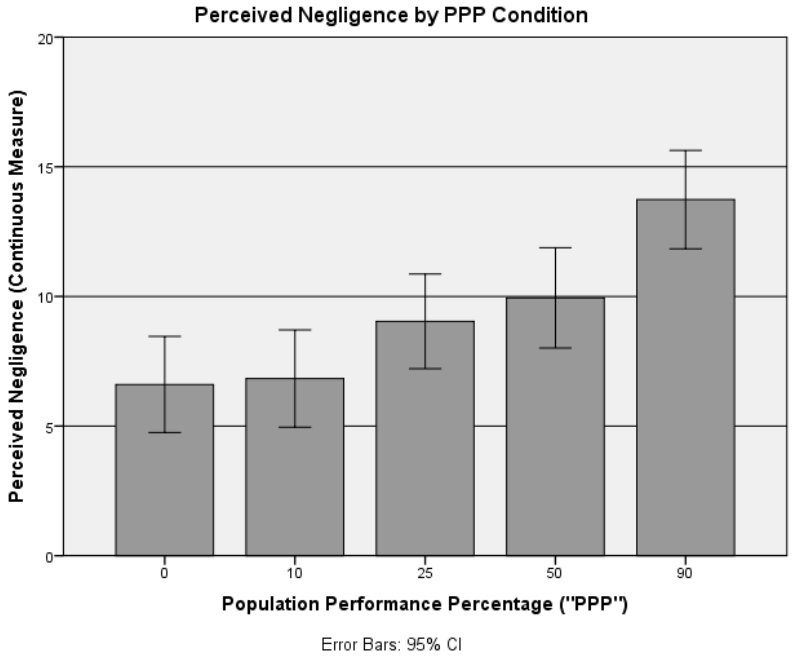


Figure 1.4. Mean negligence rating by PPP condition in Experiment Four.

An alternative analysis categorizing participants' verdict patterns provided additional support for the empirical reasonable person hypothesis. I used the same categories as in Experiment Three: verdict patterns consistent with a PPP threshold above 49% were classified as

[127] Negligence ratings in the 90% PPP condition were greater than those in the 50% PPP condition (M = 9.943, SD = 7.020), t(52) = 2.997, Bonferroni-corrected p = .042, d = .412; greater than those in the 25% PPP condition (M = 9.038, SD = 6.622), t(52) = 3.252, Bonferroni-corrected p = .020, d = .447; greater than those in the 10% PPP condition (M = 6.830, SD = 6.804), t(52) = 4.897, Bonferroni-corrected p < .001, d = .673; and greater than those in the 0% PPP condition (M = 6.604, SD = 6.721), t(52) = 5.546, Bonferroni-corrected p < .001, d = .762.

reflecting an empirical reasonable person hypothesis, [128] and verdict patterns consistent with a PPP threshold between 0% and 10% were classified as reflecting an aspirational reasonable person standard.[129] The verdict patterns of 18 of 53 participants were consistent with a pure empirical standard, versus four of 53 participants whose verdicts were consistent with a pure aspirational standard. Among participants in these two categories, verdict patterns were more likely to reflect an empirical standard, $\chi^2(1\ df) = 8.909$, $p = .003$. Table 1.5 shows how participants' verdict patterns were distributed in Experiment Four.

Verdict Pattern	# of Participants
Aspirational: 0%-10% Tipping Point	1
Aspirational: 10%-25% Tipping Point	3
Empirical: 25%-50% Tipping Point	10
Empirical: 50%-90% Tipping Point	8
Always Rule for Plaintiff	1
Always Rule for Defendant	2
Other	28

Table 1.5. Participants' verdict patterns in Experiment Four.

[128] This category included two of the 32 possible verdict patterns: (1) participants who found the defendant was not negligent in the 0%, 10%, and 25% PPP conditions, but was negligent in the 50% or 90% PPP conditions; and (2) participants who found the defendant was not negligent in the 0%, 10%, 25%, or 50% PPP conditions, but was negligent in the 90% condition.

[129] This category also included two of the 32 possible verdict patterns: (1) participants who found the defendant was not negligent in the 0% PPP condition but was negligent in all other PPP conditions; and (2) participants who found the defendant was not negligent in the 0% and 10% PPP conditions, but was negligent in all other PPP conditions.

It is also noteworthy that, unlike the three previous experiments, participants in Experiment Four were not more likely than chance to rule for the plaintiff. While it is important not to place too much interpretive weight on a null result, it should be noted that participants' response patterns showed substantial pro-plaintiff bias in the three experiments involving corporate defendants, but no such bias in the one study involving individual defendants.[130]

Overall, these findings demonstrate that participants were influenced by PPP, and suggest that participants' tipping point was more consistent with the empirical reasonable person hypothesis (which posits a tipping point at or around 50%) than with the aspirational reasonable person hypothesis (which posits a tipping point at or around 10%). Similar to previous experiments, however, participants' responses are clearly influenced by factors beyond PPP alone.

b. Follow-Up Analyses on the PPP-Negligence Relationship

As in previous experiments, I examined a subset of the data comprised of (only) participants' first verdicts. With limited statistical power due to 53 participants being spread over 5 conditions, I did not observe any effect of PPP on participants' first negligence ratings, $F(4, 48) = 1.058$, $p = .387$. Descriptively, however, participants' first negligence ratings were highest in the 90% PPP condition ($M = 10.643$, $SD = 8.073$; all other conditions M's ≤ 8.600). A comparison of participants' first negligence ratings in the 90% PPP condition and the average of participants negligence ratings across the other four PPP conditions ($M = 6.410$, $SD = 7.639$) was nearly significant, $t(51) = 1.753$, $p = .086$, with a Cohen's d of 0.546.

[130] *See* note 76, *supra*.

In addition, because the cognitive faculty at the heart of each negligence case was different in this experiment, I was interested in whether my primary findings were driven by one or two specific cases, or whether the relationship between PPP and negligence ratings was relatively stable across cases. The former might indicate that participants vary their conception of the reasonable man standard across cognitive capacities, perhaps applying an empirical standard to vision but an aspirational standard to decision making. The latter would suggest that participants generally use an empirical standard across cognitive characteristics.

Table 1.6 summarizes, in terms of descriptive statistics, participants' negligence ratings by case vignette and PPP condition. For four of the five cases (all but the reaction time case), negligence ratings were highest in the 90% PPP condition. In all five cases, participants rated the defendant as more negligent when PPP was 90% than when PPP was 0% or 10%.

Case vignette (by relevant cognitive characteristic)	0% PPP	10% PPP	25% PPP	50% PPP	90% PPP
Decision making	3.364 (5.372) N = 11	3.100 (3.872) N = 10	6.875 (7.136) N = 16	12.000 (9.539) N = 3	14.000 (8.042) N = 13
Vision	8.154 (9.036) N = 13	3.455 (5.922) N = 11	6.200 (4.940) N = 10	10.813 (7.609) N= 16	13.667 (9.452) N = 3
Memory	14.667 (6.658) N = 3	14.769 (7.429) N = 13	13.000 (8.062) N = 11	15.700 (6.913) N = 10	19.375 (2.446) N = 16
Hearing	7.125 (7.805) N = 16	12.000 (9.539) N = 3	9.615 (7.911) N = 13	5.818 (7.427) N = 11	15.800 (6.070) N = 10
Reaction Time	2.600 (5.082) N = 10	2.625 (4.395) N = 16	8.667 (10.970) N = 3	7.077 (7.279) N = 13	5.636 (6.845) N = 11

Table 1.6. Mean negligence rating by PPP condition for each of the five case vignettes. Standard deviations are included in parentheses.

While the results for each vignette generally resemble the overall findings reported in Figure 1.4, there are not enough responses for each combination of vignette and PPP condition to reliably conduct, for each individual case vignette, the same statistical analyses that I conducted using the data set as a whole. (For instance, only three participants responded specifically to the vision case vignette in the 90% PPP condition.) As a proxy, I examined vignettes individually by grouping together participants' verdicts where less than half of the population would have succeeded where the defendant failed (i.e. the 0%, 10%, 25% PPP conditions, collectively "low

PPP conditions") with those where half or more of the population would have succeeded (the 50% and 90% PPP conditions, collectively "high PPP conditions"), then conducted t tests to evaluate whether those groups' negligence ratings differed. The results revealed that participants found defendants significantly more negligent in high PPP conditions than low PPP conditions for the vision case ($t(51) = 2.475$, $p = .017$, $d = .709$), the memory case ($t(51) = 2.267$, $p = .028$, $d = .623$), the decision making case ($t(51) = 4.410$, $p < .001$, $d = 1.320$), and nearly so for the reaction time case ($t(51) = 1.846$, $p = .071$, $d = .509$). There was no such difference for the hearing case, $t(51) = .871$, $p = .388$, $d = .245$. In sum, it does not seem that the relationship between PPP and negligence verdicts is being driven by one or two outlying scenarios. Rather, participants seem to rate defendants as more negligent at higher levels of PPP (and especially at 90% PPP) across cases.

c. Responses to the Most People and Reasonable Person Questions

Similar to the previous experiments, participants' negligence verdicts, their answers to the reasonable person question, and their answers to the "most people" question agreed in the vast majority of cases (74%). Among the other cases, participants' negligence verdicts were descriptively more likely to agree with their response to the "most people" question (6.8% of cases) than with their response to the "reasonable person" question (5.3% of cases).[131] Responses to these questions do not suggest any meaningful number of participants employed an aspirational reasonable person standard in Experiment Four.

[131] Interestingly, in the other 37 instances of disagreement, participants' responses to the "most people" and "reasonably prudent and careful person" questions agreed with one another, but disagreed with their negligence verdict. It is difficult to interpret what happened in these instances from the data at hand. Perhaps participants were not satisfied as to the causal relationship between the defendant's cognitive failure and the plaintiff's injury, or perhaps broader or more abstract concerns (e.g. a sense that the plaintiff should recover) led participants to deviate from the outcome suggested by their reasonable person assessments.

4. The Effect of Scenario on Negligence Ratings

The case vignettes in this experiment were more heterogeneous than those used in the first three experiments, as each involved the failure of a different cognitive faculty. With this in mind, it should be noted that participants' negligence verdicts varied significantly across case vignettes, $F(4, 208) = 18.485$, $p < .001$. *Post hoc* comparisons among the different case vignettes revealed that participants rated the defendant as more negligent in the memory case vignette ($M = 15.962$, $SD = 6.546$) than in any of the other four,[132] that participants rated the defendant as more negligent in the hearing vignette than the reaction time vignette,[133] and that verdicts did not vary significantly across any other pairs of case vignettes.

IV. Discussion and Future Directions

The findings reported in this Chapter help clarify how lay decision makers define the reasonable person. Experiments One and Two compared the influences of positive information (about whether most people would have done what the defendant did under the circumstances) and economic information (about whether precautions were cost-justified) on negligence verdicts. I found that positive information consistently influenced how participants assessed negligence, indicating that the reasonable person standard, as applied, is at least in part a positive standard. In contrast, economic information had no influence on participants' negligence evaluations.

[132] Negligence ratings for the memory case exceeded those for the vision case ($M = 7.925$, $SD = 7.696$), $t(52) = 7.197$, $p < .001$, $d = .989$, for the hearing case ($M = 9.377$, $SD = 8.068$), $t(52) = 4.734$, $p < .001$, $d = .650$, for the reaction time case ($M = 4.679$, $SD = 6.375$), $t(52) = 8.295$, $p < .001$, $d = 1.139$, and for the decision making case ($M = 7.472$, $SD = 7.775$), $t(52) = 5.759$, $p < .001$, $d = .790$ (all p values Bonferroni corrected).

[133] $t(52) = 3.671$, Bonferroni-corrected $p = .006$, $d = .504$.

Experiments Three and Four sought to disentangle two plausible positive interpretations of the reasonable person standard: empirical and aspirational. Experiment Three, which involved corporate defendants considering but declining to take certain precautions, produced mixed evidence. Experiment Four, which involved individual defendants who exhibited a variety of cognitive failures, produced clear evidence in support of the empirical interpretation.

Cumulatively, my findings indicate that, when applying the reasonable person standard, lay decision makers apply something closer to the empirical standard than an aspirational or economic one. In one sense, this may be intuitive. As discussed in Part II, the plain language surrounding the reasonable person standard is consistent with an empirical interpretation. Jury instructions speak to what an *ordinarily* prudent person would do—to what is common in one's community. Thus, mock jurors who rely on positive information about human behavior are, on some level, simply following directions.

In another sense, however, these findings are quite surprising: they conflict with common scholarly assertions and assumptions concerning how legal decision makers define the reasonable person. As intuitive as the empirical reasonable person standard may be, it has not been especially popular in legal scholarship. Indeed, in the absence of data on how decision makers operationalize the standard, legal scholars have frequently assumed that it is operationalized in aspirational terms.[134] Moreover, the empirical view has not fared especially well in the ongoing debate about how the reasonable person *should* be interpreted. Indeed, prominent scholars have asserted that

[134] *See, e.g.*, Rachlinski, *supra* note 1, at 1058 ("law defines the reasonable person in idealized terms rather than in terms consistent with actual behavior"); HERBERT, *supra* note 38, at 12 (describing the reasonable person as "devoid, in short of any human weakness, but odious character who stands like a monument in our Courts of Justice, vainly appealing to his fellow-citizens to order their lives after his own example.").

an empirical reasonable person standard is an *impossibility*, as any statistical attempt to define the reasonable person cannot succeed.[135]

While the findings reported in this Chapter provide support for the empirical reasonable person hypothesis, it is important to avoid drawing broad conclusions too quickly. While positive information consistently influenced decision makers' negligence verdicts in my studies, it was not by any means dispositive. If it were, then no participants would have excused a defendant when 90% of similarly-situated people would have taken the relevant precaution. The reasonable person standard is undoubtedly *more* than a purely empirical standard. Negligence is a complex construct, and one potentially interesting avenue for future research would be to continue identifying factors that influence verdicts and the relative weight those factors have on participant decisions. Researchers might ultimately strive to develop a mathematical or cognitive model of negligence decisions.

Relatedly, one limitation of Experiments Three and Four is that the positive information participants reviewed was limited to five specific conditions: participants were told that either 0%, 10%, 25%, 50%, or 90% of others would have succeeded where the defendant failed. While these conditions were chosen because of their relevance to the economic and empirical hypotheses,[136] they were not evenly spaced along a number line. The present studies leave open the possibility that the relation between positive information and perceived negligence is linear. Future research might investigate the *shape* of the function describing the relation between positive information and negligence ratings, perhaps spacing PPP values at 10% intervals from 0% to 100%. Doing so

[135] Miller & Perry, *supra* note 2.
[136] I know of no theoretical perspective on the reasonable person that would predict differences between 51% PPP and 75% PPP, for example, should affect participants' judgments.

could provide a more nuanced account of how participants treat positive information and, therefore, how they conceptualize the reasonable person.

Another interesting future direction would be to manipulate the jury instructions given to participants. As noted above, the standard instructions given to participants in my experiments are arguably more consistent with a positive interpretation than an economic interpretation. But what if participants were given economically-oriented instructions? Such instructions could be modeled on the Restatement (Third) of Torts. If participants receiving such instructions shift their interpretation of the reasonable person to apply a more economic standard, then the findings would highlight a path forward for advocates of an economic standard. However, if participants' verdicts are unaffected by change in instructions, it might highlight a deeper problem with the economic view of the standard. It may be that most people simply do not (and cannot) understand the concepts of reasonableness or negligence in economic terms—particularly if the exercise of defining the reasonable person entails calling upon conceptions of the "generalized other" or otherwise drawing on experience.

Future research can also explore the role that motivated cognition[137] plays in lay decision makers' construction of the reasonable person standard. This Chapter has largely assumed that lay constructions of the reasonable person are constant. However, it could be that lay decision makers are flexible in their operationalization of the reasonable person standard, and may (consciously or unconsciously) apply a different standard if needed to help them reach a desired outcome. One avenue for exploring this could be to continue investigating whether and how the standard is conceptualized differently in cases involving corporate defendants than in cases

[137] Scholarship on motivated reasoning suggests that the interpretation of legal standards can be unconsciously shaped by preferred results. *See, e.g.,* Avani Mehta Sood, *Cognitive Cleaning: Experimental Psychology and the Exclusionary Rule*, 103 GEO. L.J. 1543 (2014).

involving individual defendants. Although there are already a number of documented findings indicating that decision makers are biased against corporate defendants in tort cases, no work to my knowledge has documented whether that bias is related to an underlying shift in how the reasonableness standard is operationalized.

If future research indicates that the relationship between positive information and negligence verdicts is robust, it would set the stage for important future work at the intersection of law and psychology. Prior research in cognitive psychology demonstrates that people often misjudge their own and others' tendencies and cognitive abilities. For instance, people often misestimate what generic others can see[138] and what they know,[139] which can compound the hindsight biases often at play in legal situations.[140] If legal decision makers endeavor to deploy an empirical reasonable person standard, but systematically *misjudge* others' cognitive capacities (as this cognitive psychology research suggests), then the decision makers' verdicts may be systematically biased. Future research can explore the contours of this bias—and means for debiasing—in tort law and in other areas of law that incorporate the reasonable person standard.

---

[138] *See, e.g.*, Daniel T. Levin & Bonnie Angelone, *The Visual Metacognition Questionnaire: A Measure of Intuitions About Vision*, 121 AM. J. PSYCHOL. 451 (2008).
[139] *See, e.g.*, Raymond S. Nickerson, *The Projective Way of Knowing: A Useful Heuristic That Sometimes Misleads*, 10 CURRENT DIRECTIONS IN PSYCH. SCIENCE 168 (2001); Joachim Krueger & Russell W. Clement, *The Truly False Consensus Effect: An Ineradicable and Egocentric Bias in Social Perception*, 67 J. PERSONALITY & SOCIAL PSYCH. 596 (1994).
[140] *See, e.g.*, Jeffrey J. Rachlinski, *A Positive Psychological Theory of Juging in Hindsight*, 65 U. CHI. L. REV. 571 (1997); Erin M. Harley, *Hindsight Bias in Legal Decision Making*, 25 SOCIAL COGNITION 48 (2007).

CHAPTER 2

REPRESENTING TECHNOLOGICAL "MINDS": ANTHROPOMORPHIZING TECHNOLOGY INFLUENCES POLICY OPINIONS AND LEGAL DECISIONS

Abstract

Research indicates that anthropomorphizing autonomous agents can affect people's decision about those agents. But much of the research investigating this impact treats anthropomorphism as a unitary, "hair-triggered" inference. An alternative view is that anthropomorphism is multi-dimensional, involving various types of tacit and explicit inferences that may or may not be "hair-triggered" and can vary independently. In this Chapter, I present three studies testing this alternative view in one important contemporary context: legal decisions about self-driving cars. My results demonstrate that people are reluctant to broadly anthropomorphize self-driving cars in a legal context, but that specific anthropomorphic attributions they do make are predictive of their policy opinions (e.g. should self-driving cars be allowed on the roads in their state?) and their legal judgments (e.g. should the manufacturer be held liable for an accident?). Further, my findings support a multidimensional view of anthropomorphism, in which some, but not all, types of human thought are likely to be attributed to machines. These different types of attributions may be uncorrelated, or only loosely correlated, and can have distinct (and sometimes opposing) influence on participants' decisions.

Introduction

On the evening of March 18, 2018, an Uber self-driving vehicle struck and killed Elaine Herzberg, a pedestrian crossing the street in Tempe, Arizona (Burkitt, 2018). The accident became a point of national discussion as the public began grappling with some of the thorny issues surrounding self-driving cars. Are we comfortable sharing the roads with autonomous machines? Should Arizona have allowed Uber to test self-driving vehicles on state roadways? Was Uber legally responsible for the accident?

Imagine that each of these questions is directly posed to you. What sorts of factors might affect your decision? You will likely be curious about how far away from Ms. Herzberg the self-driving car was when Ms. Herzberg entered the roadway, and whether Ms. Herzberg looked before stepping into the road. You may also want to know what the road conditions were like, and whether the car swerved or otherwise responded to the pedestrian. More fundamentally, however, your answers to all of these questions will likely be shaped by your understanding (or lack of understanding) about how self-driving cars work.

Philosophers and scientists have long noted the peculiar human tendency to "see human" in the world around us—to anthropomorphize entities that clearly are not human (see Heider & Simmel, 1944; Barrett & Keil, 1996; Gray, Gray, & Wegner, 2007; Epley, Waytz, & Cacioppo, 2007). People dress their pets in human-like clothing, curse at cars when they do not start, and think that video games are cheating them. Research suggests that the tendency to anthropomorphize is a stable individual trait (Waytz, Cacioppo, & Epley, 2010). People are especially prone to anthropomorphize gadgets such as self-driving cars, where the workings of those gadgets are not well understood or their behavior appears unpredictable (Epley, Akalis, Waytz, & Cacioppo, 2008; Waytz, Morewedge, et al., 2010; Bartz, Tchalova, & Fenerci, 2016). And the effects of conceptualizing devices as human-like might have consequences in the

external world: there is evidence that anthropomorphizing gadgets can influence, for example, purchasing (Yuan & Dennis, 2017) and ethical (Bartneck & Hu, 2008) decisions about machines.

Surprisingly little research, however, has examined the potential role of anthropomorphism in the critical law and policy issues surrounding autonomous machines. A careful examination of this relation is needed. Autonomously-functioning machines are cruising our streets, writing our news articles, and even caring for our sick. The increasing role of these machines is raising pressing legal questions—questions that call for answers soon (see, *e.g.*, Richards & Smart, 2016; Calo, 2015; Jaeger & Levin, 2016; Shank & DeSanti, 2018; Lin, Abney, & Bekey, 2011). This Chapter investigates whether and how the answers to these questions might be influenced by anthropomorphism. It also uses the salient legal context surrounding self-driving cars as a window to examine the cognitive processes underlying anthropomorphism and their relation to decision-making more broadly.

Conceptualizing Autonomous Machines: Two Competing Views

A rich psychological literature examines how people conceptualize ambiguous agents like autonomous machines (e.g. Epley, Akalis, Waytz, & Cacioppo, 2008; Waytz, Morewedge, et al., 2010; Kahn et al., 2012; Levin, Killingsworth, et al., 2013). In this literature, the dominant theoretical account is that people are quick to anthropomorphize the unfamiliar (*see* Epley, Waytz, & Cacioppo, 2007; Epley, Akalis, et al., 2008; Waytz, Morewedge, et al., 2010; Bartz, Tchalova, & Fenerci, 2016). Anthropomorphism is described as "hair-triggered," meaning that people have a default, reflexive tendency to conceive of ambiguous agents as humanlike, using their own experience as "an automatic base for induction" (Epley, Waytz, & Cacioppo, 2007). Conceiving of ambiguous agents as human versus nonhuman has "a powerful impact on whether

those agents are treated as moral agents worthy of respect and concern or treated merely as objects, on how people expect those agents to behave in the future, and on people's interpretations of these agents' behavior in the present" (ibid). People can, and sometimes will, overcome or "correct" their default to anthropomorphize, but doing so requires mental effort and, therefore, motivation.[141]

I refer to this account as the "promiscuous anthropomorphism account" because it suggests a general tendency to (over-)anthropomorphize the unknown. Thus, the account is analogous to ideas in the developmental literature of "promiscuous normativity"—the tendency of young children often spontaneously infer or attribute social norms in novel situations where there actually are none (Schmidt, Butler, Heinz, & Tomasello, 2016)—and of "promiscuous teleology"—the tendency of children to broadly understand objects and entities as designed for purposes, even when they are not (Keleman, 1999).

Two features of the promiscuous anthropomorphism account merit special attention. First, on this account, all instances of anthropomorphism—from describing a self-driving car as making a decision to arguing that self-driving cars have consciousness—arise from the same cognitive process. People anthropomorphize quickly and globally, in response to minimal cues, and pare back anthropomorphic attributions only with deeper thought and effort (Epley, Waytz, & Cacioppo, 2007; Epley, Waytz, Akalis, & Cacioppo, 2008). Second, and relatedly, the promiscuous anthropomorphism account generally treats anthropomorphism as a single, unidimensional construct. Different anthropomorphic attributions reflect different points on the continuum of a single anthropomorphism construct, on which people exhibit "the same kind of

[141] Epley, Waytz, and Cacioppo (2007) identify several factors that can provide or reduce this motivation: knowledge about the relevant agent, sociality/loneliness, and effectance (i.e. the need to feel effective in one's environment).

variability [in strength] that occurs in the strength of any attitude" (Epley, Waytz, Akalis, & Cacioppo, 2008, p. 145).

On the promiscuous anthropomorphism view, even minimal and relatively superficial cues are enough to prompt people to—at least initially—equate machines and humans. Some studies cited in support of this hypothesis demonstrate that people "mindlessly" apply social categories and norms in interactions with technology. For example, Nass, Moon, and Green (1997) found that when people interacting with male-voiced or female-voiced computers tend to gender-stereotype the computers as if they are people. Similar experiments demonstrate that people apply racial categories and show in-group/out-group effects in computer interactions (Nass, Isbister, & Lee, 2000; Nass, Fogg, & Moon, 1996). Other research reveals that participants "mindlessly" apply social norms like politeness and reciprocity in computer interactions (Fogg & Nass, 1997; Nass, Moon, & Carney, 1999).

Additional evidence that minimal cues can elicit anthropomorphism comes from studies that ask participants to directly rate the "humanness" of machines on various dimensions. For instance, Siino, Chung, & Hinds (2008) found that participants who interacted with a robot during sessions in which it made human-like disclosures about its feelings rated the robot as more likeable than participants who interacted with a robot in sessions where it made only task-relevant disclosures, and that this difference was mediated by participants' responses to a short Likert scale asking how "human" the robot was across several dimensions. Nass and his colleagues demonstrated that elements of a computer's "personality" influenced participants' subsequent questionnaire ratings about the computer's intelligence and about their attitude toward the computer (Moon & Nass, 1996; Nass, Moon, Fogg, Reeves, & Dryer, 1995).

DiSalvo, Gemperle, Forlizzi, and Kiesler (2002) found that certain external design features of robot heads lead to increased perceptions of humanness.

While these examples all involved interactions with machines, research has shown that people do not have to see or interact with technology to anthropomorphize it: in some cases, participants attribute agency to technology based merely on framing cues in short descriptions. Several studies have presented participants with one- to three-sentence descriptions of technological devices framed either anthropomorphically or non-anthropomorphically (Bartz, Tchalova, & Fenerci, 2016; Waytz, Morewedge, et al., 2010; Epley, Akalis, Waytz, & Cacioppo, 2008). The anthropomorphic framings use agentive language (e.g. "When Emotoboard senses your typing style…") and label the device's functions as unpredictable (Waytz, Morewedge, et al., 2010). The non-anthropomorphic framings use non-anthropomorphic language (e.g. "You can program Emotoboard such that it will ...") and do not reference how predictably it functions. On subsequent questionnaires, participants tended to ascribe more human-like qualities to the devices that were framed anthropomorphically. Collectively, these studies suggest that indirect evidence (e.g. testimony) about technology can prompt people to attribute agency to it. They also provide evidence for the view that anthropomorphism is hair-triggered, as simple framing cues prompted participants to make more anthropomorphic attributions.

Evaluating this evidence requires consideration of the anthropomorphism questionnaires used in these studies. Most of the questionnaires are versions of a measure developed by Epley and colleagues (Epley, Akalis, Waytz, & Cacioppo, 2008; Waytz, Cacioppo, & Epley, 2010; May & Monga, 2013; Bartz, Tchalova, & Fenerci, 2016). On Epley and colleagues' measure, participants use a Likert scale to rate their agreement with five statements about an agent: the agent has "a mind of its own," "has free will," "has consciousness," "has intentions," and "can

experience emotions." These items were selected on the grounds that they "reflect properties captured in previously used measures of attribution of human uniqueness and higher order cognition to human targets" (Waytz, Cacioppo, & Epley, 2010). Participants' responses to these items tend to be highly correlated within particular classes of agents (ibid.). As a result, when researchers have limited time to ask participants questions (e.g. when participants are in an fMRI scanner), researchers may use only one statement, such as the "mind of its own" statement (e.g., Waytz, Morewedge, et al., 2010).

Two features of this approach to measuring anthropomorphism merit discussion. First, the questions posed by Epley and colleagues' measure are quite broad. That is, the questions are not narrowly tailored to the agent they are asked about or to the human-like cognitive skills that are relevant in a particular scenario. Thus, these questions test whether participants make what I refer to as "far-transfer attributions"—attributions that require extrapolation far beyond what was presented in the stimuli. For a participant to agree that a technology they read about in a vignette has "consciousness," they must generalize from specific behaviors described in text to attribute broader, subjective feelings and qualities typically associated with humans (Baker, Hymel, & Levin, 2018). "Far-transfer attributions" contrast with "near-transfer attributions," which are more narrowly tailored inferences drawn about specific cognitive skills or abilities. "Near-transfer attributions" acknowledge agents' abilities to do particular tasks often associated with having a mind (such as choosing among options, remembering information, or knowing things), while not necessarily entailing broader attributions (such as consciousness or feelings) (ibid.).

Second, and relatedly, the Epley measure of anthropomorphism is deliberately unidimensional. All questions are intended to capture the same construct. Anthropomorphism is

ultimately scored by combining participants' responses to all of the questions, and, as noted above, participants' responses to those questions are correlated.

These features of Epley and colleagues' anthropomorphism measure make sense when viewed from a promiscuous anthropomorphism perspective. But there are other perspectives in the anthropomorphism literature. A number of recent findings appear inconsistent with the assumption that anthropomorphism is a default; in many situations, anthropomorphism seems to be slow, effortful, and selective. These findings tend to occur where researchers investigate participants' anthropomorphic attributions in the context of tasks that entail deeper consideration of, or formal reasoning about, technological entities (e.g. Fussell, Kiesler, Setlock, & Yew, 2008; Levin, Saylor, & Lynn, 2012; Jaeger & Levin, 2016).

For example, Levin, Killingsworth, Saylor, Gordon, and Kawamura (2013) asked participants to consider three types of agents (a computer, a robot, and a human) and make specific behavioral predictions for each in two types of scenarios (relating to goal-directed action or object categorization). Across four experiments, participants' predictions reflected (1) a sharp, consistent distinction between the computer and the human, and (2) a difficult-to-overcome default toward conceptualizing the robot as computer-like rather than human-like. Describing the robot in anthropomorphic language (e.g. naming it "OSCAR" and describing it as having goals) did not cause participants to differentiate the robot from the computer. Neither did showing them video of "OSCAR" walking, running, and letting people pass it. It was only when participants were asked to watch and memorize OSCAR's preferential gaze that they began making more human-like behavioral predictions. Collectively, these experiments show that participants intuitively contrast the cognitive processes of humans with the cognitive processes of machines, and that—at least in this context—it takes more than anthropomorphic framing to

overcome that intuitive contrast (Levin, Killingsworth, & Saylor, 2008). This intuitive contrast can persist even after direct interactions with a robot (Levin, Harriott, Paul, Zhang, & Adams, 2013). It has also been observed in middle school students following extensive interaction with a teachable agent learning system (Hymel, Levin, Barrett, Saylor, & Biswas, 2011; Jaeger, Hymel, Levin, Biswas, Paul, & Kinnebrew, 2019).

These findings could arguably be reconciled with the promiscuous anthropomorphism account if participants were initially anthropomorphizing the relevant technologies and subsequently paring back those attributions with deeper thought. But research has suggested precisely the opposite sequence of events. Levin, Saylor, and Lynn (2012) found that participants who responded to behavioral prediction scenarios faster were actually *less likely* to anthropomorphize machines than those who responded more slowly.

This "selective anthropomorphism" pattern of results is not unique to studies using behavioral prediction paradigms. Additional studies show that people are, in some cases, reluctant to broadly anthropomorphize on scales similar to the Waytz, Cacioppo, & Epley (2010) scale (see Baker, Hymel, & Levin, 2018). Further, when people do anthropomorphize, there is evidence that it is not an all-or-nothing process. In their seminal paper, Gray, Gray, and Wegner (2007) found evidence of two independent dimensions of anthropomorphic attributions, which they labeled "experience" (the capacity for subjective experiences, such as emotions; linked to moral patiency) and "agency" (the capacity to plan and do things; linked to moral agency and responsibility).[142] Participants' attributions to a given entity often differed across dimensions; for example, participants attributed a robot a fair degree of agency, yet almost no capacity for

[142] Epley and colleagues' anthropomorphism measure includes concepts related to both agency (e.g. "has intentions") and experience (e.g. "can experience emotions"), and thus may not capture a unitary construct.

experience. More recent research further supports the idea that people make narrow and targeted anthropomorphic attributions, specific to particular features or cognitive skills, rather than broadly conceptualizing of agents as human-like across the board (Baker, Hymel, & Levin, 2018; Levin, Killingsworth, et al., 2013; Levin, Saylor, & Lynn, 2012; Jaeger & Levin, 2017). Thus, the construct of "anthropomorphism" might be best understood as multi-dimensional, covering multiple types of tacit and explicit inferences that may or may not be hair-triggered (see generally Gray, Gray, & Wegner, 2007; Baker, Hymel, & Levin, 2016).

Finally, there is an interesting piece of circumstantial evidence that supports the idea that at least some forms of anthropomorphic attribution are not hair-triggered. Specifically, when people *do* make anthropomorphic attributions to machines, they are sometimes mediated by cognitive conflict (Baker et al., in prep). This suggests that—on at least some dimensions—people are not anthropomorphizing by default but instead doing so in an effort to resolve inconsistencies between their observations and their existing knowledge (see Jaeger & Levin, 2016). For example, people may only anthropomorphize a machine if they observe it engaging in behavior that is inconsistent with specific expectations (*e.g.*, when a robot displays preferential looking patterns, as in Levin, Killingsworth, et al., 2013).

To sum up, then, there is evidence that in some situations people are biased to equate man and machine, while in other situations people are biased to strongly distinguish them. One way of thinking about what differentiates these situations is to consider the degree of deep consideration they invite. The selective anthropomorphism pattern tends to emerge when people are engaged in *meaning making*—that is, when people thinking deeply to try to integrate information about a machine within their existing knowledge structure. People might engage in meaning making because they are explicitly asked to reason about a technological agent (ibid.).

But they might also engage in meaning making because the agent is embedded in the context of a broader, consequential decision that requires careful consideration (Jaeger & Levin, 2016).

The Role of Anthropomorphism in Decision Making

Researchers are increasingly exploring the role that anthropomorphism plays in various types of technology-related decisions (e.g., Aggarwal & McGill, 2007; Yuan & Dennis, 2017; de Melo, Gratch, & Carvenale, 2015; Jaeger & Levin, 2016). Findings indicate that the relation between anthropomorphism and decision making varies depending on the *type* of decision being made.

A core operating assumption of the psychological literature on decision making is that decision making results from two distinct processes or systems (e.g.,Sloman, 1996; Kahneman & Frederick, 2002; Loewenstein, O'Donoghue, & Bhatia, 2015). The first is typically fast, unconscious, and effortless, often relying on heuristics to produce results. The second is typically slow, conscious, controlled, and deliberative. Researchers have generally observed more anthropomorphism when embedding potentially-anthropomorphized entities in the context of quick, intuitive, system-one judgments (e.g. how much they like a product; Aggarwal & McGill, 2007) than in decision-making contexts designed to elicit deeper reflection (de Melo, Gratch, & Carvenale, 2015; Bartneck & Hu, 2008).

This dichotomy is illustrated by an experiment conducted by Fussell, Kiesler, Setlock, and Yew (2008). Participants were asked to read a series of vignettes describing interactions between a medical patient and either a human or a robot health interviewer. After reading the vignettes, participants did two things. First, participants responded true or false to statements about the traits of the interviewer (which included human mental traits) in a speeded decision-

making paradigm (participants were instructed to respond as quickly as possible). Second, participants responded yes or no to items on a 16-item survey designed to elicit careful thought about robot properties (e.g. "If a robot acts happy today, is it likely to act happy tomorrow?"; "Can a robot imagine things it has not learned?"). The researchers found that participants anthropomorphized robot health interviewers on the first measure—responding "true" to human mental characteristics as often and as rapidly for the robot interviewer as for the human interviewer—but anthropomorphized robots less on the survey designed to elicit careful thought.

These findings demonstrate that decision making context can affect the degree to which participants anthropomorphize machines. But perhaps more interesting is the reverse relation—the ways that anthropomorphism affects people's decisions. It has been suggested that "perceptions of mind in autonomous agents can have a profound impact on people's decision making" (de Melo, Gratch, & Carnevale, 2014). And recent work suggests that anthropomorphism can influence decisions in various domains. For example, in economics, participants engaged in an ultimatum game with computerized agents respond differently to agents they perceive as more anthropomorphic, offering more money to agents that express more emotion (de Melo, Gratch, and Carnevale, 2014). In marketing, there is evidence that participants evaluate products they perceive as anthropomorphic more favorably (Aggarwal & McGill, 2017) and pay more for them (Yuan & Dennis, 2017).

Waytz, Heafner, and Epley (2014) studied how anthropomorphic attributions to self-driving cars affect trust in them. Participants experienced a physical driving simulator in one of three conditions: normal, in which participants controlled the car themselves; agentic, in which the car functioned autonomously; and anthropomorphic, in which the agentic car also had a name, a gender, and a voice. At the end of the simulation, participants experienced a "virtually

unavoidable" accident that was clearly caused by another driver in the simulation. Waytz, Heafner, and Epely found: (1) that participants made more anthropomorphic attributions to the car in the anthropomorphic condition than the agentic condition, and more in the agentic condition than in the normal condition; (2) that participants in the anthropomorphic condition trusted the vehicle more than participants in the agentic condition, and more in the agentic condition than in the normal condition; and (3) that participants blamed their own car for the accident less in the anthropomorphic condition than the agentic condition (while blaming the car least in the normal, non-self-driving condition). Thus, there is evidence that anthropomorphic attributions affect trust in vehicles and can affect moral concepts like blame.

Anthropomorphism also influences decision-making in utilitarian moral dilemmas like the trolley problem. For example, people's responses about various agents on measures of anthropomorphic attributions are correlated with their choices about whether to "save" those agents in dilemmas (Strait, Briggs, & Scheutz, 2013). Relatedly, participants are more likely to "save" humans in utilitarian moral dilemmas when they have been primed to consider that person's mental states (i.e. thoughts and feelings) than when they have not (Majdandžić, Bauer, Windischberger, Moser, Engl, & Lamm, 2012).

One area where anthropomorphism might have particularly important consequences is the law. The growing number of sophisticated technologies that we interact with raise important legal questions, which are being taken up in a growing literature on robotics law and policy (e.g., Calo, 2015; Richards & Smart, 2016; Vladeck, 2014). It seems that the legal system's responses to these questions will likely be influenced by the anthropomorphic attributions made (or not made) by the relevant decision makers (for review, see Jaeger & Levin, 2016).

In prior work, Jaeger and Levin (2017) evaluated the degree to which participants anthropomorphized particular technologies predicted their verdicts in mock negligence cases involving various autonomous technologies. In a series of three studies, participants acted as mock jurors, deciding a case in which a plaintiff was injured by an autonomously-functioning machine (a self-driving car, an autonomous drone, or a robotic nurse) owned and operated by the defendant. After reading the case summary, participants decided whether the defendant was liable for the plaintiff's injury, then completed two additional measures: a six-item measure of cognitive conflict (Levin, Adams, et al., 2013) and a six-item anthropomorphism measure. The anthropomorphism consisted of two items from Epley and colleagues' questionnaire probing "far-transfer" attributions (asking the extent to which the device "has consciousness" and "has emotions") and four items probing "near-transfer" attributions (asking the extent to which the device "makes decisions," "uses strategies," "considers alternatives," and "knows things") (Epley, Akalis, Waytz, & Cacioppo, 2008; Baker, Hymel, & Levin, 2018).

Across all three studies, Jaeger and Levin (2017) found that participants were selective with their anthropomorphic attributions: near-transfer and far-transfer attributions were uncorrelated, with participants generally making near-transfer attributions while resisting far-transfer attributions. Further, across all three studies, Jaeger and Levin found significant relations between anthropomorphic attributions and negligence verdicts against the car's manufacturer. Importantly, the direction of the relation varied from near-transfer attributions to far-transfer attributions. Near-transfer attributions were consistently *negatively* correlated with finding the manufacturer negligent, while far-transfer attributions were consistently *positively* correlated with finding the manufacturer negligent. These results are similar to those published by Baker, Hymel, and Levin (2018), who reported than near-transfer and far-transfer attributions

had opposite effects on memory for short stories about technological agents. Further, Jaeger and Levin (2017) found, in all three studies, that cognitive conflict was positively correlated with far-transfer attributions. This suggests that participants did not make far-transfer attributions by default but instead made them to resolve inconsistencies in their understandings of autonomous machines (Baker et al., in prep).

Figure 2.1 is a path diagram summarizing the relations among variables in the first of Jaeger and Levin's three studies. Jaeger and Levin's second and third studies partially replicate the pattern observed in this path diagram. Specifically, cognitive conflict positively predicted far-transfer attributions in all three studies, far-transfer attributions positively predicted findings of manufacturer negligence in all three studies, and near-transfer attributions negatively predicted findings of manufacturer negligence in two of three studies.

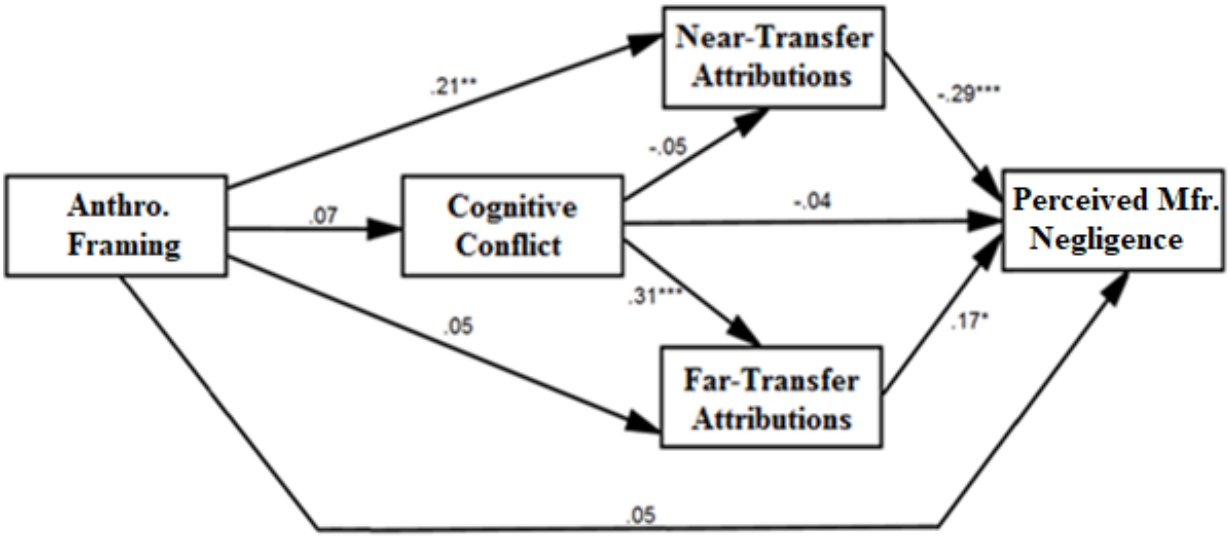


Figure 2.1. Path diagram illustrating relations among variables in Jaeger and Levin (in prep). Specifically, the diagram illustrates relations among framing (anthropomorphic versus non-anthropomorphic), cognitive conflict, near-transfer attributions, far-transfer attributions, and findings of manufacturer negligence. All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

The studies presented in this Chapter expand on Jaeger and Levin (2017), investigating the influences of cognitive conflict, near-transfer anthropomorphism, and far-transfer anthropomorphism on a wider array of law and policy questions. These studies build on the existing literature in four important ways.

First, prior work has rarely specified the causal relations between anthropomorphic framing (i.e. depicting or describing a machine as human-like), anthropomorphic attributions (i.e. participants indicating an inference or belief that the machine has human-like qualities), and decisions about machines. Building on the model in Figure 2.1 above, in this Chapter I specify

and test one possible model: framing affects anthropomorphic attributions, which in turn affect decisions.

Second, while many researchers have alluded to anthropomorphism playing an important role in real-world decisions, the evidence of such a role is limited (cf. Waytz, Heafner, & Epley, 2014). More often than not, anthropomorphism is studied in artificial contexts where there is no broader decision to be made beyond whether the relevant gadget or agent is human-like (e.g. Epley, Akalis, Waytz, & Cacioppo, 2008). Where there is a decision to be made, the relevant information available to participants is constrained and disproportionately focused on anthropomorphism (e.g. Yuan & Dennis, 2017). The present studies embed the relevant agent (a self-driving car) in a broader decisional context, probing the extent to which anthropomorphism influences decisions when it is just one among many decision-relevant factors.

Third, and relatedly, the idea that anthropomorphism is hair-triggered (Epley, Waytz, & Cacioppo, 2007) is largely based on studies that lack broader decision-making context. In such studies, researchers often find that framing manipulations, such as describing the agent in anthropomorphic language, trigger attributions of agency. But other studies reviewed above suggest that anthropomorphism may be less hair-triggered in more consequential contexts. Thus, in addition to looking at how anthropomorphic attributions affect broader decisions, the present studies examine whether the relation between framing and anthropomorphic attributions persists in the face of such broader decisions.

Fourth, much of the research reviewed above treats anthropomorphism as a single construct (Yuan & Dennis, 2017; Strait, Briggs, & Scheutz, 2014), though other research suggests the broad label of "anthropomorphism" may encompass multiple conceptually and practically distinct forms of inferences (e.g. Gray, Gray, & Wegner, 2007; Baker, Hymel, &

Levin, 2018). Examining anthropomorphism in context of broader legal decisions allows me to not only test whether different aspects or forms of anthropomorphism are correlated with one another, but also whether they have different influences on participants' downstream decisions.

Experiment 1

In Experiment 1, participants were asked to read a news article about the Elaine Herzberg accident described at the beginning of this Chapter, then answer a series of five law- and policy-oriented questions about self-driving cars. Subsequently, participants completed a six-item measure assessing their anthropomorphic attributions to the self-driving car, and a six-item cognitive conflict scale.

Building on Jaeger and Levin (2017), I had four key hypotheses. First, I expected that near-transfer and far-transfer anthropomorphic attributions to the self-driving car would be uncorrelated, with participants generally resisting far-transfer attributions. Second, I expected that cognitive conflict would be positively correlated with far-transfer attributions to the self-driving car. Third, I expected that participants' near-transfer and far-transfer attributions would predict whether they found the self-driving car's manufacturer negligent in a hypothetical lawsuit, with near-transfer attributions functioning as negative predictors and far-transfer attributions as positive predictors (consistent with prior studies). Fourth, and relatedly, I expected that near-transfer attributions would predict favorable policy opinions about self-driving cars, while far-transfer attributions would be negatively predictive.

In addition to investigating these relations, I also manipulated whether participants received additional information pertaining to the Herzberg accident. Many details related to ensuing investigation were reported in the days following the initial news story, and I wanted to investigate the extent to which the relations of interest varied with an expanded set of facts. Of

particular interest: some initial investigation reports suggested that the accident was effectively unavoidable once Herzberg walked in the roadway—that no human driver could have stopped the vehicle in time.[143] I expected that such information would result in more legal judgments favoring the self-driving car's manufacturer, but I wanted to investigate whether it might also lead to broader adjustments of policy attitudes toward self-driving cars (as this information could might place greater attention on the relation between the car's capabilities and human capabilities, perhaps inviting increased anthropomorphism of the car).

Method

Participants. Eighty participants completed the experiment through Amazon Mechanical Turk, an online platform frequently used by researchers to recruit participants. Due to the possibility that some participants may attempt to complete the study more than once, an *a priori* exclusion rule was established: participants would be excluded if (i) they had the same Mechanical Turk ID as another participant, (ii) their response was associated with the same IP address as another participant, or (iii) they responded from the same geographic coordinates as another participant. Pursuant to this rule, two responses were excluded based on IP address redundancies.

The final sample of 78 participants included 50 men and 28 women, ranging in age from 19 years to 69 years with an average age of 36.08 years. The participants were divided between two conditions, defined below: the control condition (N = 37) and the "Additional Facts"

[143] It should be noted that many of the investigation's initial conclusions were subsequently called into question.

condition (N = 41). All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedure. Upon giving consent, participants were informed via written instructions that they would be reading a recent news article involving technology, then asked questions about the article and the technology described in it.

Participants then read a March 20, 2018 CNN Money article titled "Uber's self-driving car killed someone. What happened?" The article, which was 358 words in length, laid out initial details of the March 18, 2018 accident in which a self-driving Uber Vovlo XC90 SUV struck and killed Elaine Herzberg in Tempe, Arizona. The article noted that an Uber employee was in the vehicle at the time of the accident, stated that Arizona's governor had recently updated an executive order allowing self-driving cars on state roads with or without test drivers behind the wheel, and noted that other companies including Google/Waymo, GM, and Intel were testing self-driving vehicles in Arizona. A full copy of the article is included in the Appendix B. To ensure participants read the article, participants were prompted to write two sentences summarizing what they found to be the most interesting points.

Next, approximately one-half of participants—those randomly assigned to the "Additional Facts" condition—were told that they would receive additional information about the Herzberg accident and that, for the purposes of the study, they should assume the information was accurate. These participants were then told that investigation of the accident revealed that there was no time for the vehicle to stop in response to the pedestrian entering the roadway, that no human driver could have possibly responded fast enough to avoid colliding with the pedestrian, and that the accident was "unavoidable" once the pedestrian entered the roadway (see Appendix B for "Additional Facts" prompt). The other half of participants—those in the

"Control" condition—did not review any additional facts. Aside from this difference, the study was identical for participants in both conditions.

After reading about the Herzberg accident, participants rated their agreement with a series of four statements about self-driving cars on a scale from 0 (completely disagree) to 100 (completely agree). The statements were: (1) "I would be comfortable riding in a self-driving car", (2) "The city and state where I live should allow self-driving cars on the roads, with drivers behind the wheel as a safeguard", (3) "The city and state where I live should allow self-driving cars on the roads, with **or without** drivers behind the wheel as a safeguard", and (4) "I am comfortable with the idea of sharing the roads with self-driving cars." Participants' responses to these four questions were highly correlated with one another (all *r*'s >= .67), so responses were averaged into one variable—"policy opinions"—for the purpose of all analyses.

Following these four questions, participants were asked to imagine that Ms. Herzberg's estate had filed a negligence suit against Uber. After receiving representative jury instructions, participants responded on a six-point scale ranging from 1 = "No. I am very confident the defendant was not negligent." to 6 = "Yes. I am very confident the defendant was negligent."

After rendering their verdicts, participants completed a six-item anthropomorphism scale about the Uber self-driving car described in the news article. For each item, participants rated their agreement with statements about the self-driving car on a seven-point Likert scale from 1 = "Completely Disagree" to 7 = "Completely Agree." As in Jaeger and Levin (2017), two items on the anthropomorphism scale probed "far-transfer" anthropomorphic attributions: "the car has consciousness" and "the car has emotions." The other four items probed "near-transfer" anthropomorphic attributions: "the car makes decisions," "the car uses strategies," "the car considers alternatives," and "the car knows things."

Next, participants completed a six-item cognitive conflict measure adapted from Levin, Adams, et al. (2013). This scale also asked participants to rate their agreement with six statements on a seven-point Likert scale ranging from 1 = "Completely Disagree" to 7 = "Completely Agree." The six statements were, "when considering my opinions and responses to items in this study…": (1) "I was always certain about my opinions" (reverse scored), (2) "If I were allowed to, I would go back and change some of my opinions," (3) "At times I worried that some of my opinions were inconsistent with my other opinions," (4) "I **never** had difficulty putting together facts to form my opinions" (reverse scored), (5) "Some of my opinions were inconsistent with my previous beliefs," and (6) "I was uncomfortable with some of my opinions."

Finally, participants completed basic attention check questions (all participants passed), answered questions about their prior familiarity with the Herzberg case, and answered basic demographic questions.

Results and Discussion

Control Condition versus Additional Facts Condition. Participants in the Control (N = 37) and Additional Facts (N = 41) conditions did not differ significantly in their policy opinions (Control Mean = 45.47; Additional Facts Mean = 48.91; $t(76) = .524$, $p = .602$), their near-transfer attributions to the Uber self-driving car (Control Mean = 4.81; Additional Facts Mean = 5.09; $t(76) = .881$, $p = .381$), their far-transfer attributions to the Uber self-driving car (Control Mean = 1.99; Additional Facts Mean = 2.39; $t(76) = 1.134$, $p = .261$), or their cognitive conflict (Control Mean = 3.32; Additional Facts Mean = 3.12; $t(76) = -.710$, $p = .480$). Participants in the Control condition were significantly more likely to find that Uber acted

negligently (M = 3.97) than participants in the Additional Facts condition (M = 2.80), as expected given that the additional facts were exculpatory for Uber, t(76) = -3.45, p = .001.

Near-Transfer versus Far-Transfer Anthropomorphic Attributions. Because the Control versus Additional Facts manipulation did not significantly affect participants' anthropomorphic attributions, I combined data from the two conditions to compare participants' near-transfer attributions to their far-transfer attributions. Participants were far more likely to make near-transfer attributions to the Uber self-driving car (M = 4.96) than far-transfer attributions (M = 2.199), t(77)=11.65, p<.001. Further, participants' near-transfer attributions were not significantly correlated with their far-transfer attributions (r = .016, p = .887).[144]

Path Analyses. In order to evaluate the relations among variables, I conducted two path analyses. The first path analysis, presented in Figure 2.2, examined the influences of condition (Control versus Additional Facts), cognitive conflict, near-transfer attributions, and far-transfer attributions on participants' policy opinions (R^2 = .23). The second path analysis, presented in Figure 2.3, examined the influences of condition (Control versus Additional Facts), cognitive conflict, near-transfer attributions, and far-transfer attributions on participants' verdicts in a hypothetical negligence case against Uber (R^2 = .36).

[144] The same pattern of results is obtained for each condition, analyzed separately.

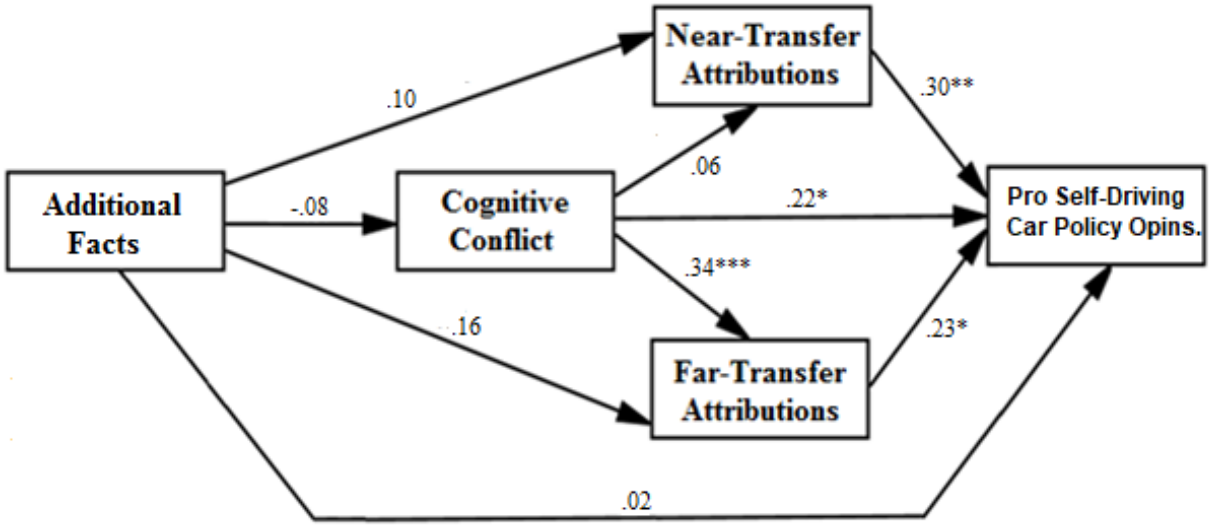


Figure 2.2. Path diagram illustrating relations among condition (Control versus Additional Facts), cognitive conflict, near-transfer attributions, far-transfer attributions, and favorable policy opinions toward self-driving cars in Experiment One. All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

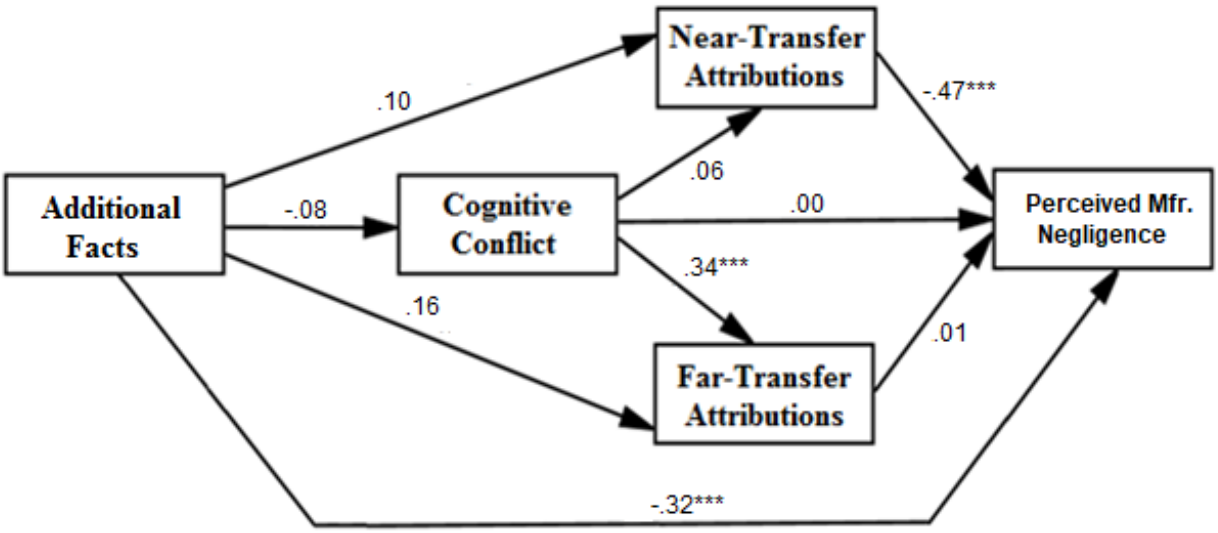


Figure 2.3. Path diagram illustrating relations among condition (Control versus Additional Facts), cognitive conflict, near-transfer attributions, far-transfer attributions, and findings that Uber was negligent in a hypothetical lawsuit in Experiment One. All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

As shown in Figure 2.2, anthropomorphic attributions to the Uber self-driving car predicted favorable policy opinions about self-driving cars more generally (echoing Waytz, Heafner, & Epley, 2014). Interestingly—and contrary to our hypothesis—both near-transfer and far-transfer attributions were *positively* correlated with favorable policy opinions. In other words, participants who conceived of a self-driving car as making decisions and having strategies generally had more favorable opinions toward self-driving cars, but participants who conceived of a self-driving car as having consciousness and emotions *also* had more favorable opinions toward self-driving cars. In light of the fact, noted above, that near-transfer attributions and far-transfer attributions were uncorrelated, it seems the two types of attributions provide independent, additive bases for supporting self-driving cars.

In Figure 2.3, in contrast, only near-transfer attributions were related to participants' negligence verdicts. The relation was strong and negative: the more participants believed that self-driving cars make decisions and have strategies, the less likely they were to fault Uber for the accident. This negative relation is consistent with the predictions in Jaeger and Levin (2016) that anthropomorphic attributions "may tend to exculpate the owner or manufacturer," either because the malfunctions of technologies that make their own decisions and execute their own strategies are less foreseeable to defendants, or because the technologies themselves act as additional agents in the scenario and effectively "soak up" some of the blame. Far-transfer attributions, however, were unrelated to negligence verdicts in this study.

The significant positive relation between cognitive conflict and far-transfer attributions in both Figure 2.2 and Figure 2.3 is also noteworthy. This relation was also observed in all three studies in Jaeger and Levin (2017), and analogous relations have been observed in other paradigms (Baker et al., in prep).

In sum, Experiment 1 demonstrated that participants' anthropomorphic attributions are significant predictors of their policy opinions about self-driving cars and their negligence verdicts in a hypothetical case involving a self-driving car accident. Further, Experiment 1 provided additional evidence that anthropomorphism is a multi-dimensional construct: near-transfer and far-transfer anthropomorphic attributions were uncorrelated in the study, with each having unique influences on policy opinions and negligence verdicts. Finally, Experiment 1 arguably yielded evidence that people are quite selective with some types of anthropomorphic attributions: participants rarely made far-transfer attributions, and the far-transfer attributions they made were associated with cognitive conflict, indicating that such attributions are more likely products of mental struggle than "hair-triggered" defaults.

Experiment 1 does, however, have its limitations. First, because the condition manipulation did not affect participants' anthropomorphic attributions, the causal inferences that can be drawn from the findings are limited. It is arguably just as likely that participants revised their anthropomorphic attributions based on their policy opinions and verdicts rather than forming policy opinions and verdicts based on their anthropomorphic attributions (see Simon, 2004). This is particularly true because, in this Experiment, participants gave their policy opinions and verdicts *before* completing the anthropomorphism scale.

Second, and relatedly, while the cognitive conflict measure in Experiment 1 was phrased so as to concern the responses to items in the study, it was only administered toward the end of the study. Without a baseline measure at the beginning of the study, it is difficult to tell whether the cognitive conflict reported by participants reflected situational responses to the study materials or broader, more dispositional tendencies to experience cognitive conflict. This weakens the argument that the observed relation between cognitive conflict and far-transfer attribution is evidence that participants are initially reluctant to make far-transfer attributions.

Third, the effect of near-transfer attributions on policy opinions and negligence verdicts in Experiment 1 could conceivably be explained without any reference to anthropomorphism at all. Participants might be glossing over the anthropomorphic language in the near-transfer prompts and instead answering based on their beliefs about how well the car functions in a purely mechanical sense. In other words, many participants might respond to "the car makes decisions" as if the prompt were "the car's software computes responses to environmental stimuli," providing an undifferentiated assessment of how well the car functions. On this view, the relation between near-transfer attributions and favorable policy opinions toward self-driving cars would be utterly unsurprising: people who think self-driving cars function better are more

inclined to think that self-driving cars should be on the road. Similarly, the relation between near-transfer attributions and negligence verdicts might simply reflect participants' judgments that manufacturers of better-functioning cars are generally less negligent than manufacturers of worse-functioning cars.

Experiment 2 was designed to address these three limitations. The first limitation was addressed by changing the condition manipulation and re-sequencing participants' tasks. With respect to the manipulation, I dropped the "additional information" manipulation and instead manipulated story framing. Some participants read versions of a news story edited to describe the self-driving car in anthropomorphic terms, while others read versions of a news story edited to describe the self-driving car in mechanical terms. I expected that anthropomorphic framing might lead participants to make more anthropomorphic attributions and, ultimately, influence their legal and policy opinions. With respect to task sequence, participants in Experiment 2 completed the anthropomorphism scale *before* being asked about policy opinions or a hypothetical lawsuit, mitigating the possibility that responses to the latter questions is driving anthropomorphism. The second limitation, related to the cognitive conflict measure, was addressed by incorporating a baseline measure of cognitive conflict at the outset of the study before any information about self-driving cars was introduced. The third limitation was addressed by the addition of questions about self-driving cars' *mechanical* functioning, which were controlled for in analyses of anthropomorphism.

Experiment 2

Method

Participants. 158 participants completed the experiment through Amazon Mechanical Turk. The same exclusion rule used in Experiment 1 was applied again in Experiment 2, excluding eight responses. An additional three responses were excluded because the participants failed basic attention checks.

The final sample of 147 participants included 102 men, 44 women, and 1 participant who preferred not to identify. The participants were divided between two conditions, defined below: the Anthropomorphic Framing condition (N = 72) and the Mechanical Framing condition (N = 75). The sample ranged in age from 20 years to 68 years with an average age of 34.26 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedure. Upon giving consent and reviewing general instructions, participants began by completing a baseline measure of cognitive conflict. This measure was identical to the one described in Experiment 1, but phrased to ask about their experiences over the past week rather than their experiences during the study (e.g. "When considering my decisions over the past week, I have always been certain about my decisions").

Participants were then randomly assigned to read one of two articles: the CNN Money article used as the stimulus in Experiment 1 or an edited version of a 2017 ABA Journal article by Steven Seidenberg titled "Who's to Blame When Self-Driving Cars Crash?" The Seidenberg (2017) article describes a self-driving car accident that killed a man named Joshua Brown. Brown was riding in a Tesla set in "autopilot" mode when the Tesla collided with a tractor trailer, which Tesla's systems apparently failed to detect. The Seidenberg article was edited

down to a length of approximately 300 words to approximate the length of the CNN Money article. The second article was added to allow stimulus rotation, to ensure that the findings of Experiment 1 were not an artifact of the specifics of the CNN Money article.

Importantly, participants in Experiment 2 could see one of two versions of the assigned article. Participants in the "Anthropomorphic Framing" condition viewed a version of the article that was edited to include extensive anthropomorphic framing (e.g. the vehicle was given a name, such as "Sporty," and descriptions included phrases such as "Sporty did not pay attention to the pedestrian" and "Sporty was making all of its own decisions"). Participants in the "Mechanical Framing" condition viewed a version of the article that was edited to include extensive mechanical framing (e.g. the vehicle was referred to exclusively as an "Uber Volvo XC90 SUV," and descriptions included phrases such as "The car's processor was processing input from other sensors at the time" and "the car was being navigated entirely by its onboard computers.")

Thus, participants in Experiment 2 read one of four possible article variants, reviewing either the CNN or ABA article in either the Anthropomorphic Framing or Mechanical Framing condition. There were no "Additional Facts" provided after the initial article in Experiment 2.

After reading the assigned article, participants completed the same set of questionnaires they completed in Experiment 1, but in a different sequence and with one addition. As noted above, participants in Experiment 2 completed the six-item anthropomorphism scale after reading the article, and *before* being asked to share policy opinions or make legal judgments. Further, immediately after completing the anthropomorphism scale, participants responded to two questions about the mechanical operations of self-driving cars in general: (1) "Overall, how effectively do you think self-driving cars can obtain and process information from the

environment?" and (2) "Overall, how effectively do you think self-driving cars can use the information they obtain to produce correct and safe actions?" These questions about mechanical quality were intended to provide controls for analyses involving anthropomorphism. Only after completing these questions did participants respond to the policy opinion and lawsuit questions.

At the end of the study, participants completed a variant of the six-item cognitive conflict questionnaire used in Experiment 1, but phrased such that the items concerned participants' responses to items in the study. Participants' score on the initial baseline cognitive conflict questionnaire were subtracted from their scores on this second, study-focused cognitive conflict questionnaire, giving a more accurate measurement of "situational cognitive conflict" arising from the study materials.

Results and Discussion

As expected, participants' responses did not differ based on whether participants read the CNN Money or the ABA article for any of the dependent measures. Therefore, all statistical analyses collapse across the two articles.

Anthropomorphic Framing versus Mechanical Framing. Participants in the Anthropomorphic Framing (N = 72) and Mechanical Framing (N = 75) conditions did not differ significantly in their policy opinions (Anthropomorphic Framing Mean = 48.68; Mechanical Framing Mean = 50.33; $t(145) = .343$, $p = .732$), their negligence verdicts (Anthropomorphic Framing Mean = 4.03; Mechanical Framing Mean = 3.77; $t(145) = -1.027$, $p=.306$), their near-transfer attributions to the self-driving car (Anthropomorphic Framing Mean = 5.10; Mechanical Framing Mean = 4.90; $t(145 = -.921$, $p = .359$), their far-transfer attributions to the self-driving car (Anthropomorphic Framing Mean = 2.53; Mechanical Framing Mean = 2.11; $t(137.015) = -$

1.46, p = .147), their attributions of mechanical quality to the self-driving car (Anthropomorphic Framing Mean = 63.61; Mechanical Framing Mean = 61.23; t(145) = 0.609, p = .543), or their situational cognitive conflict[145] (Anthropomorphic Framing Mean = -.54; Mechanical Framing Mean = -.57; t(145) = -.173, p = .863). In sum, the framing manipulation did not have any significant effect on participants' responses. This may be because individual differences in participants' tendency to anthropomorphize—which has been observed to be a stable individual trait (Waytz, Cacioppo, & Epley, 2010)—are swamping any effect of situational factors specific to the study.

Anthropomorphic Attributions and Perceptions of Mechanical Quality. Participants were far more likely to make near-transfer attributions to self-driving cars (M = 5.00) than far-transfer attributions (M = 2.32), t(146)=15.372, p<.001. Also consistent with prior findings, participants' near-transfer attributions were not significantly correlated with their far-transfer attributions (r = .037, p = .658). Participants' near-transfer attributions were correlated to their responses on the new mechanical quality items (r = .348, p < .001), but far-transfer attributions were not correlated with attributions of mechanical quality. Table 2.1 summarizes the bivariate correlations among these composite variables.

[145] As noted above, "situational cognitive conflict" was calculated by subtracting participants' score on the initial baseline cognitive conflict measure, which concerned their decisions in the past week, from the participants' score on the second, study-specific cognitive conflict measure.

Correlations

		Near-Transfer Attributions	Far-Transfer Attributions	Attributions of Mechanical Quality
Near-Transfer Attributions	Pearson Correlation	1	.037	.348**
	Sig. (2-tailed)		.658	.000
	N	147	147	147
Far-Transfer Attributions	Pearson Correlation	.037	1	-.077
	Sig. (2-tailed)	.658		.353
	N	147	147	147
Attributions of Mechanical Quality	Pearson Correlation	.348**	-.077	1
	Sig. (2-tailed)	.000	.353	
	N	147	147	147

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2.1. Correlation matrix summarizing relations among near-transfer attributions, far-transfer attributions, and attributions of mechanical quality in Experiment Two.

Path Analyses. Paralleling the analyses in Experiment 1, I conducted two path analyses. The first path analysis, presented in Figure 2.4, looked at the influences of condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and attributions of mechanical quality on participants' policy opinions (R^2 = .57). The second path analysis, presented in Figure 2.5, looked at the influences condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and attributions of mechanical quality on participants' verdicts in a hypothetical negligence case against the manufacturer of the self-driving car (R^2 = .15).

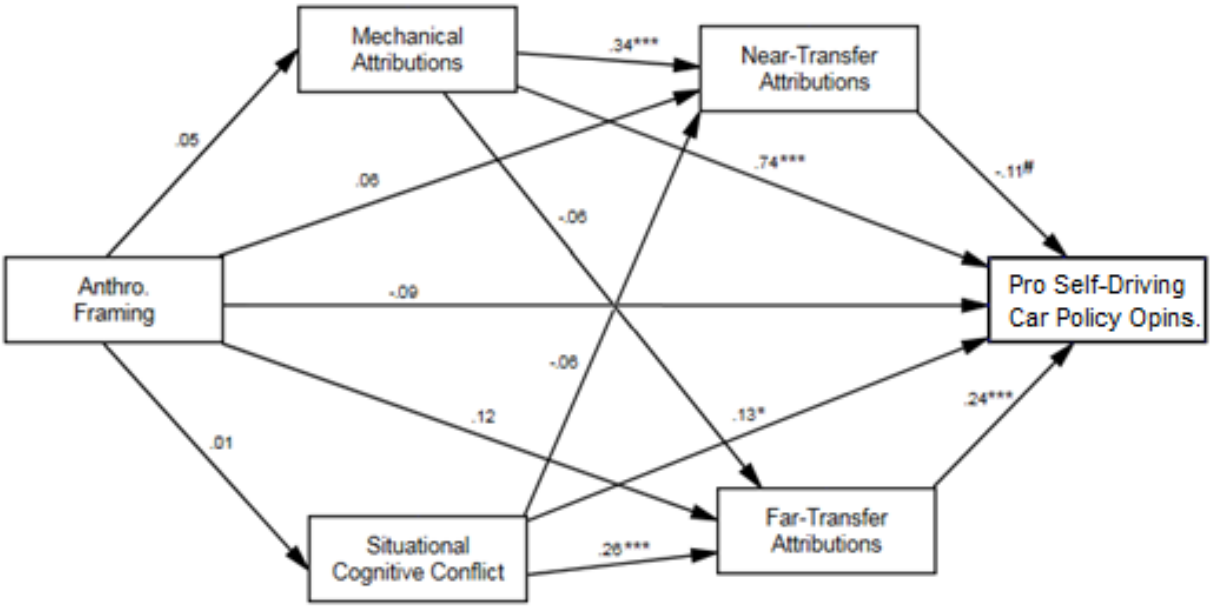


Figure 2.4. Path diagram illustrating relations among condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and favorable policy opinions toward self-driving cars in Experiment Two. All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The link marked # reflects $p = .057$. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

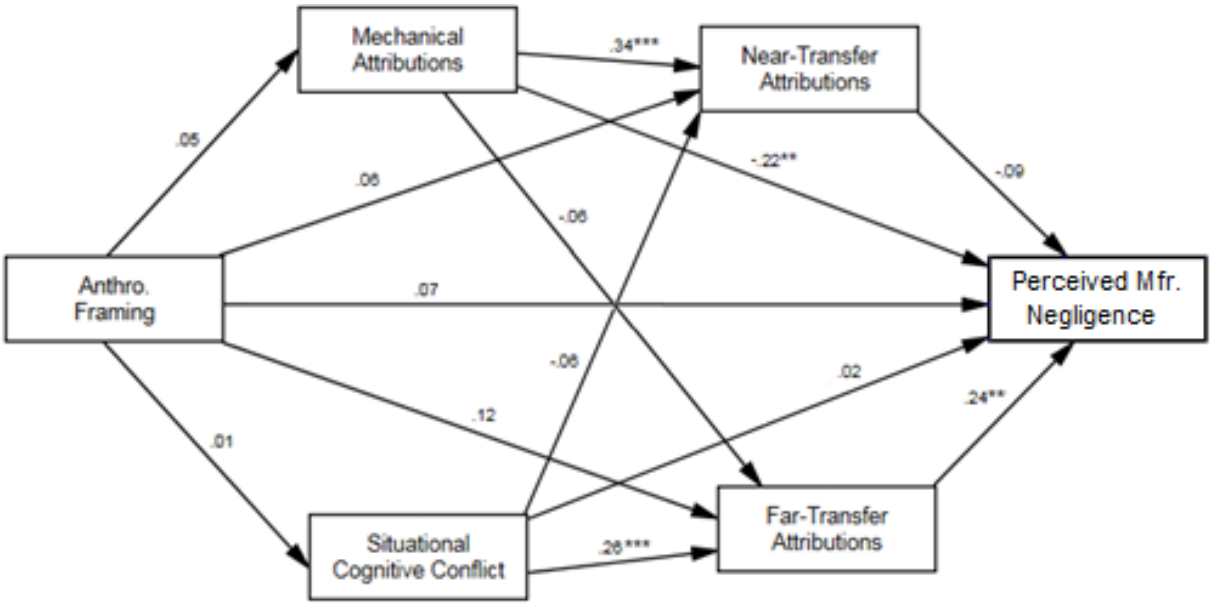


Figure 2.5. Path diagram illustrating relations among condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and findings that the manufacturer was negligent in a hypothetical lawsuit in Experiment Two. All reported path coefficients are standardized. . All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

As shown in Figure 2.4, attributions of mechanical quality to self-driving cars are, naturally, closely related to favorable policy opinions about them ($\beta = .74$, $p<.001$). However, far-transfer anthropomorphic attributions ($\beta = .24$, $p<.001$) and perhaps near-transfer anthropomorphic attributions ($\beta = -.11$, $p=.057$) also affect policy opinions above and beyond the influence of simple attributions of mechanical quality. Further, echoing Experiment 1, near-transfer and far-transfer attributions have *opposing* relations with policy opinions, underscoring again the multi-dimensional nature of anthropomorphic attributions and their affects.

As shown in Figure 2.5, a similar pattern of significant and non-significant relations arises in the context of negligence judgments. Participants' attributions of mechanical quality

were negatively related to their findings of manufacturer negligence (β = -.22, p=.007). This makes sense: the better participants believe the self-driving car functions, the less likely they are to hold the manufacturer liable. Interestingly, far-transfer attributions were a significant predictor of negligence judgments in the *positive* direction in Experiment 2 (β = .24, p = .002) while near-transfer attributions were not (β = -.09, p = .254). I defer interpretation of this finding to the General Discussion.

Finally, both Figure 2.4 and Figure 2.5 show a significant positive relation between the more-refined measure of situational cognitive conflict used in Experiment 2 and far-transfer anthropomorphic attributions. As mentioned above, this relation has been previously observed in other paradigms (Baker et al., in prep), and it has appeared in all of my previous studies of anthropomorphism in law and policy contexts (Jaeger & Levin (2017)).

In sum, Experiment 2 replicated and extended key findings of Experiment 1. First, Experiment 2 confirmed the basic point that participants' anthropomorphic attributions are significant predictors of their opinions and decisions on matters of law and policy. Second, Experiment 2 provided more evidence of the multi-dimensionality of anthropomorphic attributions, as near-transfer and far-transfer attributions were again uncorrelated with one another and tended to have opposing relations with participants' legal opinions and decisions. Third, and most importantly, Experiment 2 demonstrated that the influences of anthropomorphic attributions on participants' opinions and decisions remain significant even when controlling for participants' attributions of *mechanical quality* to the self-driving cars. In other words, anthropomorphism affects decision making in ways that go beyond participants' simple impressions about a device's mechanical functioning.

One limitation of Experiment 2 is that, as in Experiment 1, the framing manipulation did not affect participants' anthropomorphic attributions (or other dependent measures). While participants varied in the extent to which they anthropomorphized the self-driving cars, that variability was not related to the experimental manipulation. Thus, the causal direction of the relations observed in Experiment 2 are not entirely clear (though, as explained in the General Discussion, the sequence of tasks in the study suggests anthropomorphism is influencing decision-making rather than the other way around).

Prior research suggests that, in order to manipulate anthropomorphic attributions in decision making contexts, linguistic framing may not be enough. Levin, Killingsworth, Saylor, Gordon, and Kawamura (2013) similarly found that text-based framing manipulations did not create differences across conditions, but ultimately observed differences when participants in the anthropomorphic framing condition viewed videos of a robot exhibiting human-like behavior (a preferential looking pattern) and answered questions about that behavior. Similarly, I designed Experiment 3 to include a series of short video prompts and questions highlighting anthropomorphic or mechanical features of cars, with the aim of further differentiating the Anthropomorphic Framing and Mechanical Framing conditions.

Experiment 3

Method

Participants. 200 participants completed the experiment through Amazon Mechanical Turk. 20 responses were excluded using the same *a priori* exclusion rule used in Experiments 1 and 2 (one due to a repeat IP address and nineteen due to repeat geographic locations). One additional responses was excluded because the participant failed basic attention checks.

The final sample of 179 participants included 104 men, 73 women, and 2 participants who preferred not to identify. The participants were divided between two conditions, defined below: the Anthropomorphic Framing condition (N = 91) and the Mechanical Framing condition (N = 88). The sample ranged in age from 20 years to 73 years with an average age of 35.75 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedure. The stimuli and procedures were identical to those in Experiment 2, with one important exception: After completing the baseline measure of cognitive conflict, but before reading the assigned article about a self-driving car accident, participants viewed and answered questions about three short video segments (under 30 seconds). The videos and questions differed across the Anthropomorphic Framing and Mechanical Framing conditions, and were intended to supplement the manipulation.

Participants in the Anthropomorphic Framing condition viewed segments of an edited video of a driverless car engaging in a series of autonomous functions (e.g. turning right; slowing for a pedestrian; video available at https://www.youtube.com/watch?v=aaOB-ErYq6Y). The video, produced by Waymo, was shot from a camera situated in the backseat and looking out the front windshield of the car, so the viewer can see the wheel turning on its own as the car navigates the roads. After viewing each of three segments from this video, participants answered an anthropomorphically-framed question about events in that segment. For example, after watching a video segment in which a self-driving car stopped for a red light, participants were asked: "What did the car see before it stopped?" Participants' answer choices were all framed in anthropomorphic terms (e.g. "It saw a red light."), so that participants had to endorse as correct a statement anthropomorphizing the car.

Alternatively, participants in the Mechanical Framing condition viewed clips in which mechanic Scotty Kilmer provides instructions on how to change the spark plugs in a standard, non-self-driving car (video available at https://www.youtube.com/watch?v=I3qwK-eXEIc). After each segment, participants answered one question emphasizing mechanical qualities of the car. For example, after watching the initial segment on preparing to change the spark plugs, participants were asked: "The mechanic suggests changing spark plugs when the engine is cold because most engines have cylinder heads made out of _____." Participants could choose aluminum (the correct answer), copper, or graphite.

After participants viewed and answered questions about the video segments, the remainder of the study unfolded just like Experiment 2. As in Experiment 2, the "situational cognitive conflict" variable was calculated by subtracting participants' scores on the baseline cognitive conflict measure from their score on the final, study-focused cognitive conflict questionnaire.

Results and Discussion

As expected, participants' responses did not differ based on whether participants read the CNN Money or the ABA article for any of the dependent measures. Therefore, all statistical analyses collapse across the two articles.

Anthropomorphic Framing versus Mechanical Framing. Participants in the Anthropomorphic Framing (N = 91) and Mechanical Framing (N = 88) conditions did not differ significantly in their policy opinions (Anthropomorphic Framing Mean = 50.82; Mechanical Framing Mean = 44.19; $t(177) = -1.555$, $p = .122$), their negligence verdicts (Anthropomorphic Framing Mean = 3.92; Mechanical Framing Mean = 4.26; $t(177) = 1.514$, $p=.132$), their near-

transfer attributions to the self-driving car (Anthropomorphic Framing Mean = 5.10; Mechanical Framing Mean = 4.84; t(177) = -1.330, p = .185), their far-transfer attributions to the self-driving car (Anthropomorphic Framing Mean = 2.02; Mechanical Framing Mean = 2.29; t(177) = 1.151, p = .251), or their situational cognitive conflict (Anthropomorphic Framing Mean = -.82; Mechanical Framing Mean = -.85; t(177) = -0.139, p = .890).

However, in this study, participants in the Anthropomorphic Framing condition attributed better mechanical functioning to the self-driving car than participants in the Mechanical Framing condition (Anthropomorphic Framing Mean = 68.92; Mechanical Framing Mean = 59.71; t(177) = -2.569, p = . 011). Further, it should be noted that framing significantly affect responses for two of the six individual items on the anthropomorphism scale. Participants in the Anthropomorphic Framing condition were more likely to agree with the statement that the car makes decisions (Anthropomorphic Framing Mean = 5.90; Mechanical Framing Mean = 5.39; t(177) = -2.347, p = .020) and the statement that the car uses strategies (Anthropomorphic Framing Mean = 5.28; Mechanical Framing Mean = 4.77; t(177) = -2.011, p = .046).

Anthropomorphic Attributions and Perceptions of Mechanical Quality. Participants were far more likely to make near-transfer attributions to self-driving cars (M = 4.97) than far-transfer attributions (M = 2.15), t(178)=37.558, p<.001. Unlike Experiments 1 and 2, participants' near-transfer attributions were significantly correlated with their far-transfer attributions in this study (*r* = .215, p = .004). Follow-up analysis with data split by condition revealed a significant correlation among participants in the Mechanical Framing condition (*r* = .296, p = .005), but not among participants in the Anthropomorphic Framing condition (*r* = .130, p = .221).

Similar to Experiment 2, participants' near-transfer attributions were correlated to their responses concerning mechanical quality (r = .428, p < .001), but far-transfer attributions were not correlated with mechanical quality (r = .041, *ns*). Table 2.2 summarizes the bivariate correlations among these composite variables.

Correlations

		Near-Transfer Attributions	Far-Transfer Attributions	Attributions of Mechanical Quality
Near-Transfer Attributions	Pearson Correlation	1	.215**	.428**
	Sig. (2-tailed)		.004	.000
	N	179	179	179
Far-Transfer Attributions	Pearson Correlation	.215**	1	.041
	Sig. (2-tailed)	.004		.584
	N	179	179	179
Attributions of Mechanical Quality	Pearson Correlation	.428**	.041	1
	Sig. (2-tailed)	.000	.584	
	N	179	179	179

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2.2. Correlation matrix summarizing relations among near-transfer attributions, far-transfer attributions, and attributions of mechanical quality in Experiment Three.

Path Analyses. Paralleling the prior experiments, I conducted two path analyses. The first path analysis, presented in Figure 2.6, looked at the influences of condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and attributions of mechanical quality on participants' policy opinions (R^2 = .59). The second path analysis, presented in Figure 2.7, looked at the influences condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and attributions

of mechanical quality on participants' verdicts in a hypothetical negligence case against the manufacturer of the self-driving car (R^2 = .13).

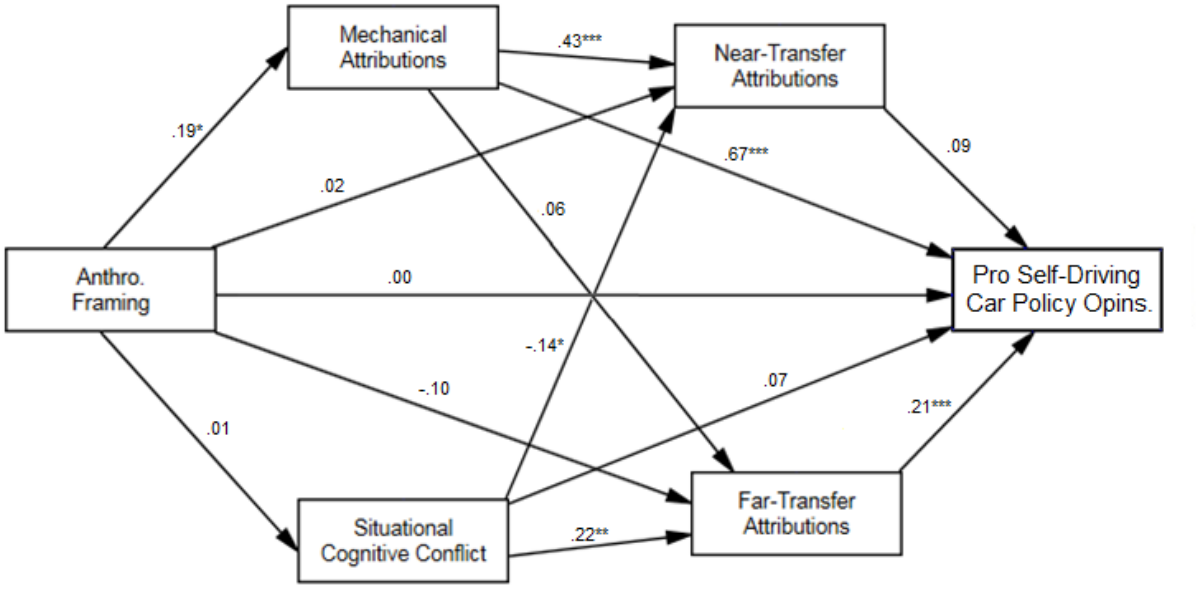


Figure 2.6. Path diagram illustrating relations among condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and favorable policy opinions toward self-driving cars in Experiment Three. All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

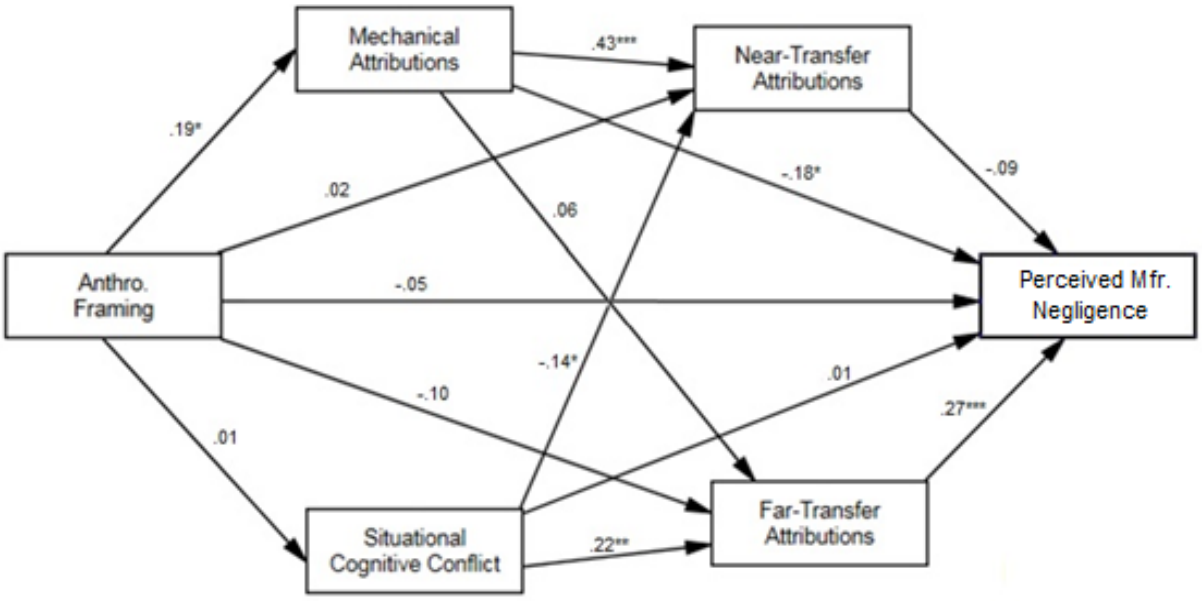


Figure 2.7. Path diagram illustrating relations among condition (Anthropomorphic Framing versus Mechanical Framing), situational cognitive conflict, near-transfer attributions, far-transfer attributions, and findings that the manufacturer was negligent in a hypothetical lawsuit in Experiment Three. All reported path coefficients are standardized. Links marked with a * are significant at the $p < .05$ level, links marked with a ** are significant at the $p < .01$ level, and links marked with a *** are significant at the $p < .001$ level. The path analysis included residuals for all endogenous variables, but these are not shown in the path model for ease of readability.

As shown in Figure 2.6, attributions of mechanical quality ($\beta = .67$, $p<.001$) and far-transfer anthropomorphic attributions ($\beta = .21$, $p<.001$) once again predicted favorable policy opinions toward self-driving cars. Near-transfer anthropomorphic attributions were not a significant predictor of policy opinions in Experiment 3 ($\beta = .09$, $p=.125$).

Figure 2.7 focuses on participants' assessment of a self-driving car manufacturer in a hypothetical lawsuit. As in Experiment 2, participants who attributed greater mechanical quality to the self-driving car were less inclined to find the manufacturer negligent ($\beta = -.18$, $p=.024$), whereas participants who made more far-transfer anthropomorphic attributions were more inclined to find the manufacturer negligent ($\beta = .27$, $p < .001$). Near-transfer attributions were unrelated to negligence decisions ($\beta = -.09$, $p = .248$).

Both Figure 2.6 and Figure 2.7 show that situational cognitive conflict was once again a significant predictor of far-transfer attributions to the self-driving car in Experiment 3 ($\beta = .22$, $p < .001$). But, unlike prior experiments, situational cognitive conflict was also negatively related with near-transfer attributions ($\beta = -.14$, $p = .04$).

Both figures show one other "new" significant relation: the framing manipulation in Experiment 3 predicted participants' attributions of mechanical quality ($\beta = .19$, $p = .011$). Participants in the Anthropomorphic Framing condition—who both read an anthropomorphically-framed story and watched video of self-driving cars in action—attributed better mechanical functionality to self-driving cars than participants in the Mechanical Framing condition. But the difference was limited to attributions of mechanical quality. Framing condition was not predictive of anthropomorphic attributions, policy opinions, or legal decisions. In hindsight, I suspect the difference in mechanical attributions arose because the videos in the Anthropomorphic Framing condition showcased a self-driving car seamlessly navigating the roads without incident, providing participants with a salient example of successful mechanical functioning.

Ultimately, Experiment 3 replicated several key findings of Experiment 2. Again, participants' anthropomorphic attributions were predictive of their policy pinions and legal decisions. And, again, anthropomorphic attributions were predictive even when controlling for participants' attributions of mechanical quality to the self-driving cars. Experiment 3 also provided further evidence of the relation between cognitive conflict and far-transfer anthropomorphic attributions.

General Discussion

The studies reported in this Chapter produced three key findings. First, people's anthropomorphic attributions to self-driving cars affected their legal decisions and policy opinions. Second, people attribute some, but not all, types of human thought to autonomous machines like self-driving cars. The different types of anthropomorphic attributions (here, near-transfer versus far-transfer) can be uncorrelated and have very different—in some cases, opposing—influences on decisions and opinions. Third, people's anthropomorphic attributions are largely unaffected by framing in contexts that prompt deep reasoning or meaning making. I will discuss each finding in turn.

In all of my experiments, I observed robust relations between participants' anthropomorphic attributions and their legal decisions, adding to a growing literature on the "profound impact" anthropomorphism can have on decision-making (e.g., de Melo, Gratch, and Carnevale, 2014; Strait, Briggs, & Scheutz, 2014). The need to understand this impact is particularly pressing in the legal context, as the increasing deployment of automated technology poses substantial challenges for our legal system (Jaeger & Levin, 2016; Calo, 2015; Richards & Smart, 2016). An understanding of how anthropomorphism influences legal decisions can inform ongoing discussions about *who* makes decisions in regulating automated technologies, and about *what evidence* they consider. Future studies could hone in on these considerations. For example, prior work suggests that robotics experts should be less likely to anthropomorphize robots than non-experts (Epley, Waytz, & Cacioppo, 2007; Levin, Adams, Saylor, & Biswas, 2013). Future research might deploy the methods used in this Chapter using samples of experts and non-experts, investigating how expertise affects anthropomorphic attributions and ultimately legal decisions. The findings of such work could contribute to the ongoing conversation about

whether the legal issues raised by automated machines are best handled by robotics experts (in the form of, for example, a federal commission) or generalist legislators and judges (Calo, 2014).

My findings also provide strong support for the proposition that anthropomorphism is multi-dimensional (Gray, Gray, & Wegner, 2007). Though often lumped into a unitary construct (e.g. Epley, Akalis, Waytz, & Cacioppo, 2008), I observed little relation between near-transfer anthropomorphism (narrow inferences about particular cognitive skills) and far-transfer anthropomorphism (broader generalizations to human experiences) in my studies. The two constructs were uncorrelated in Experiments 1 and 2, and only loosely correlated in Experiment 3, suggesting that these different anthropomorphic inferences arise from independent cognitive mechanisms. It should be noted that the concepts of near-transfer and far-transfer anthropomorphism deployed in this Chapter in some respects parallel the dimensions of "agency" and "experience" identified by Gray, Gray, and Wegner (2007). My four near-transfer prompts—"the car makes decisions," "the car uses strategies," "the car considers alternatives," and "the car knows things"—generally tap into the capacities that characterized Gray, Gray, and Wegner's "agency" dimension (i.e. the capacities to plan and do things). In contrast, my two far-transfer prompts—"the car has consciousness" and "the car has emotions"—get at the capacity for having internal subjective experiences. A similar "has consciousness" prompt was included in the Gray, Gray, and Wegner study and loaded onto their "experience" dimension. Further, Gray, Gray, and Wegner asked a number of emotion-related questions (e.g. whether the agent feels fear, or feels pleasure, or feels rage), which also loaded onto their "experience" dimension.

In addition to indicating that anthropomorphism has multiple dimensions, my studies take the additional step of demonstrating those dimensions can have diverging impacts on decision-making. To my knowledge, these experiments are the first to demonstrate this divergence. In

Experiments 2 and 3 (the experiments that accounted for attributions of mechanical quality), both attributions of mechanical quality and far-transfer attributions positively predicted support for policies favoring self-driving cars on the roads, while near-transfer attributions were not significantly predictive. However, attributions of mechanical quality and near-transfer anthropomorphic attributions made participants *less* inclined to find self-driving car manufacturers negligent after an accident, whereas far-transfer anthropomorphic attributions made participants *more* inclined to do so.

This discrepancy initially seems puzzling. But both the positive relation between far-transfer anthropomorphic attributions and negligence verdicts and the (non-quite-significant) negative relation between near-transfer attributions and negligence verdicts continue a pattern previously observed by Jaeger and Levin (in prep)—the direction of these relations is remarkably consistent across studies, though the statistical significance of these links fluctuates from study to study. Similarly, Baker, Hymel, and Levin (2018) found that near-transfer and far-transfer attributions had opposing influences on participants' memory for stories about robots. Thus, it seems these discrepancies are not anomalies, but reflect consistent operations of independent cognitive processes.

But why would anthropomorphic attributions along one dimension decrease perceptions of negligence, while attributions along another increase them? One possibility is that participants view self-driving cars as something akin to Frankenstein's monster; if the monster causes harm, they will hold Dr. Frankenstein liable for his creation. In other words, when human-like entities on the road cause harm, participants fault manufacturers for turning these entities loose on us. However, this explanation appears to be in tension with my findings concerning policy opinions. In all three experiments, far-transfer attributions were associated

with policy opinions favoring self-driving cars on the road (see also Waytz, Heafner, and Epley, 2014). Future research might investigate this tension. Perhaps people are comfortable with the idea of conscious, feeling machines on the road so long as they assume or are assured that the machines' creators are financially responsible for any harms the machines cause.

The third key finding is perhaps better characterized as a key non-finding: framing had no effect on participants' anthropomorphic attributions. While participants differed substantially in the attributions they make to self-driving cars, their differences were not affected by the language or video segments included in their stimuli. Other studies have also documented limited influence of framing on anthropomorphic attribution (for discussion, see Jaeger & Levin, 2016); these (non-)findings appear to be in tension the promiscuous anthropomorphism account, which suggests minimal cues should be enough to invite anthropomorphism (e.g. Waytz, Morewedge, et al., 2010).

I see two plausible interpretations for the minimal role of framing. First, it may be that participants' anthropomorphic attributions are primarily driven by the traits (Waytz, Cacioppo, & Epley, 2010) or knowledge (or both) that they bring with them to the study. With self-driving cars increasingly in the media and on the roads, participants may have an established conception of self-driving cars when they enter the study. The specific prompt that participants read in the study may not be enough to prompt change in broader, established concepts (Levin, Adams, Saylor, & Biswas, 2013). Second, even if participants' conceptions of self-driving cars can be influenced by the prompt presented to them, triggering such an influence might require prompts containing deep, causal challenge to participants' conceptions (e.g. Lehman, D'Mello, & Graesser, 2012). Such prompts may force participants to experience "cognitive disequilibrium," which might then lead them to more deeply analyze their understanding (Graesser, McNamara,

& VanLehn, 2005). Indeed, the consistently-observed correlation between cognitive conflict and far-transfer attribution suggests that, for the minority of participants who made far-transfer attributions, something about the stimuli challenged them to the point of experiencing disequilibrium. However, changes to surface-level features of the problem, such as whether agents are described in human-like language, may not create such a challenge for most participants.

One point that warrants careful consideration is the causal direction of the relations between anthropomorphic attributions and legal decisions/policy opinions. Neither the use of linguistic framing in the self-driving car stories (Experiments 2 and 3) nor the inclusion of video segments (Experiment 3) affected participants' responses, and making causal inferences is often difficult in the absence of effects of condition. That said, in Experiments 2 and 3, the sequence of tasks suggests that anthropomorphic attributions are influencing decisions and opinions, rather than the other way around. In those experiments, I measured anthropomorphic attributions *before* asking anything about participants' policy opinions or introducing the hypothetical lawsuit. Specifically, participants would read the relevant story, then complete the anthropomorphism scale, then respond to the four policy opinion response, and then read about and evaluate the hypothetical negligence suit. It is possible that simply reading about a self-driving car accident in the initial story led participants to consider policy opinions and liability issues related to self-driving cars, and that those considerations shaped their anthropomorphic attributions. However, it seems more likely that participants considered issues in the order they were presented. One goal of future research should be to further disentangle these possibilities. Developing a manipulation that consistently affects anthropomorphic attributions—whether with

cars or with other automated technologies (Levin, Killingsworth, et al., 2013)—could help to more clearly establish the direction of causation.

Future research can also build on the studies presented in this Chapter in other ways. It would be worthwhile to vary the hypothetical lawsuit presented to participants in two respects. First, in all of my studies, the hypothetical lawsuit was premised on a negligence claim—participants were given a specific set of negligence instructions and asked whether the manufacturer was negligent. However, a plaintiff suing a self-driving car manufacturer might proceed under theories other than negligence, and it may be that some participants who did not find the manufacturer negligent would have found the manufacturer liable under another theory. Thus, future work might have participants evaluate multiple legal claims against the manufacturer, or alternatively, ask participants in general terms whether the manufacturer should be held responsible (without specifying a legal theory). Second, participants evaluating the tort suit are likely engaging in some sort of apportionment of blame among the various actors relevant to the suit, and explicitly asking participants about this apportionment could provide another useful outcome variable. For example, participants might assign a total of 100 blame points among the actors. This might allow researchers to probe more deeply how near-transfer and far-transfer attributions affect liability judgments (e.g. do far-transfer attributions result in the car itself being blamed?)

CHAPTER 3

ACCOUNTING FOR AGENTS' MINDS: SPONTANEOUS LEVEL-2 PERSPECTIVE TAKING IS SENSITIVE TO AGENCY CUES

Abstract

Research demonstrates that people sometimes adopt others' spatial perspectives spontaneously, even when there is no communicative reason to do so. This Chapter reports a series of single-trial studies testing the degree to which such spontaneous perspective taking for schematic faces is influenced by contextual features and by cognitive load. These studies reveal that participants spontaneously adopt the perspective of schematic faces at about the same rates as observed in previous research using photographic stimuli, suggesting that perspective taking occurs for a range of abstracted agentive stimuli. They also reveal that cues designed to prompt consideration of a stimulus's agency facilitate spontaneous perspective taking, while cues designed to prompt consideration of perceptual features suppress it. Further, the studies produce mixed evidence on whether cognitive load suppresses spontaneous perspective taking, even in a single-trial design. Taken together, these findings suggest that spontaneous perspective taking in response to photographic and schematic stimuli likely reflects an overlearned social tendency that draws on domain-general cognitive resources.

Introduction

Imagine that you and a colleague are on a business lunch with a prospective client. Your colleague is seated across the table from you, and she is completely unaware of the large glob of salad dressing on her chin. When the client excuses herself to take a brief phone call, you tell your colleague about the salad dressing. She takes her napkin and starts to wipe her chin, but she is nowhere near the salad dressing—she is too far to *her* right, *your* left. You want to advise her on which way to move the napkin. Do you direct her using her perspective or your own?

In situations like this, people tend to give direction using the listener's perspective (Mainwaring, Tversky, Ohgishi, & Schiano, 2003; Schober, 1993). This is generally believed to require some additional mental effort on the part of the person giving the directions (e.g. Butterfill & Apperly, 2013; Horton & Keysar, 1996; Keysar, Barr, Balin, & Paek, 1998). But exerting this effort makes sense in context. If your goal is to communicate information that the listener can use, it is often worth paying the mental cost required to frame that information in the most useable package.

But what happens if the communication goal is removed from the equation? Imagine you are simply looking at a still picture of a woman, facing you, who is trying (and failing) to wipe the salad dressing from her chin. An experimenter (who shares your spatial perspective as you look at the picture) asks you to describe the position of the salad dressing relative to the napkin. Do you answer from your own perspective or from the perspective of the woman in the picture?

Interestingly, research demonstrates that a surprising number of people—approaching half—will answer from the perspective of the woman in the picture (Tversky & Hard, 2009; Todd et al., 2015). This is puzzling. A number of other articles suggest that perspective taking is egocentrically biased—that people default to their own perspective, and that adopting another's perspective is cognitively challenging (e.g. Epley, Keysar, Van Boven, & Gilovich,

2004; Keysar, Barr, Balin, & Brauner, 2000). While people may undertake that challenge when communicating with others (e.g. Schober, 1993), participants are not communicating with the woman in the picture. There is no obvious reason for people to frame their answers from her perspective, or even to bother calculating the woman's perspective at all.

This Chapter presents a series of five experiments that examine some of the contextual factors that prompt this sort of "reasonless" perspective taking, and explores what these factors can tell us about the cognitive bases of perspective taking more generally.

Two Perspective Taking Systems?

The foundational assumption of social cognition research is that people routinely infer others' mental states (i.e. beliefs, desires, and goals) and use those inferences in order to explain and predict behavior (Westra, 2017; Fodor 1992). This practice, referred to as "theory of mind" or "mindreading" or "perspective taking," is thought to facilitate all manners of social interaction, from coordinating a handshake (see Tversky & Hard, 2009) to carrying on a conversation (Clark & Murphy, 1982; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015; Brown-Schmidt & Heller, 2018) to establishing social institutions (Tomasello, Carpenter, Call, Behne, & Moll, 2005).

Some researchers, however, have expressed skepticism about how widespread the use of theory of mind really is in our day-to-day lives (Keysar, Lin, & Barr, 2003; Butterfill & Apperly, 2013). After all, other people's actions could be informed by an *infinite* number of beliefs, desires, and goals. What inferences do we draw about a man running toward us on the sidewalk, for example? He could be running for exercise, or because there is a swarm of bees behind him, or because he is being chased by police, or for any number of other reasons. Sorting through all

the possibilities in every social situation "would no doubt be incredibly demanding and [cognitively] effortful, and would place heavy demands on executive systems like working memory—that is, if the task were not completely intractable" (Westra, 2017, p. 4560).

Recognizing this problem, a number of psychologists have proposed two-system accounts of theory of mind (e.g. Csibra & Gergely, 1998; Leslie, 1994; Apperly & Butterfill, 2009). Here, I focus on the account of Apperly and Butterfill (2009), who suggest that people are equipped with both an "implicit" (or "minimal") theory of mind, and a more advanced, "explicit" (or "full-blown") theory of mind. The implicit system is described as working online, fast, efficient, early-developing, automatic, inflexible, and modular (meaning it is unaffected by information from other systems), but of limited capacity. In contrast, the explicit system is described as unable to function online, slow, inefficient, late-developing, effortful, flexible, and drawing on domain-general resources.

Supporters of the two-system account view findings from the field of visual perspective taking as evidence for their claims. Visual perspective taking is widely viewed as a subset of theory of mind, encompassing inferences about what other people see. In children, measures of visual perspective taking are correlated with various other measures of theory of mind (Bigelow & Dugas, 2009), including measures of belief tracking such as the "false-belief task" (Wimmer & Perner, 1983) and the "lying task" (Sodian, 1991). Further, neuroscientific research has shown a degree of overlap in regions associated with false belief tasks and visual perspective taking tasks (Aichhorn, Perner, Kronbichler, Staffen, & Ladurner, 2006).

Research in visual perspective taking has long distinguished between Level 1 perspective taking and Level 2 perspective taking (Flavell, Everett, Croft, & Flavell, 1981; Michelon & Zacks, 2006). Level 1 perspective taking entails representing *what* another person can see in a

physical space and understanding that may be different from what you can see. Level 2 perspective taking, on the other hand, entails representing *how* another person sees what that person sees, incorporating the understanding that someone viewing the same object from a different angle will see it from a different perspective (ibid.; Elekes, Varga, & Király, 2017). Using line of sight cues to infer that the person seated across from me can see (or cannot see) the book on the table in front of me is an example of Level 1 visual perspective taking. Recognizing that the print on my book appears "upside down" to the person across from me is an example of Level 2 visual perspective taking. Michelon and Zacks (2006) found that a participant's reaction time in Level 2 tasks varies with the angular disparity between the participant and the other person, whereas reaction time in Level 1 tasks does not, suggesting the two tasks entail different cognitive processes. It has subsequently been argued that Level 1 and Level 2 perspective taking reflect implicit and explicit theory of mind processes, respectively (e.g. Surtees, Samson, & Apperly, 2016).

Level 1 visual perspective taking has generally been characterized as early-developing, fast, and modular. Key evidence for this characterization comes from experiments using the "dot-perspective task" (Samson, Apperly, Braithwaite, Andrews, & Bodley-Scott, 2010; Surtees & Apperly, 2012). On a trial of the dot-perspective task, the participant views a number of dots on a computer screen, as well as an avatar that (based on line-of-sight cues) "sees" either the same number or a different number of dots than the participant. Over hundreds of trials, participants are asked to report either the number of dots visible on the screen from their own perspective ("Self" trials) or the number of dots visible from the avatar's perspective ("Other" trials). Participants tend to respond slower and make more errors when the on-screen avatar's perspective differs from their own, even on *Self* trials. This suggested that participants were

calculating the on-screen avatar's perspective online, automatically, even when it was irrelevant to the participants' task, resulting in "altercentric interference" with the participants' responses. More recent work provides evidence that this "altercentric interference" cannot be attributed to simple spatial cueing alone (Baker, Levin, & Saylor, 2016). These findings have been interpreted as evidence that Level 1 perspective taking is an instantiation of implicit theory of mind (Surtees, Apperly, & Samson, 2013; Butterfill & Apperly, 2013).

In contrast, Level 2 visual perspective taking has generally been understood as late-developing,[146] slow, and a drain on domain-general cognitive resources (Flavell, Everett, Croft, & Flavell, 1981; Michelon & Zacks, 2006; Surtees, Butterfill, & Apperly, 2012). Initial research adapting the dot-perspective task for Level 2 perspective taking did not find significant altercentric interference. Specifically, Surtees, Butterfill, and Apperly (2012) adapted the dot-perspective task into a "number perspective task," in which the on-screen avatar sat opposite an on-screen table from the participant. In each trial of the task, a number appeared on the table (either 0, 6, 8, or 9). Critically, the number could either appear the same from both the participant's and the avatar's perspective—as in the case of 0 and 8—or could appear to be a different number from each perspective—as in the case of 6 and 9. Over hundreds of trials, participants were asked to judge either how the number appeared from the avatar's perspective ("Other" trials) or how the number appeared from their own perspective ("Self" trials). In their 2012 study, Surtees and colleagues found no altercentric interference in the number perspective task. They interpreted this to mean that participants did not automatically engage in Level 2

[146] Children generally do not engage in Level 2 perspective taking until they are old enough to pass standard false belief tasks (Flavell et al., 1981).

perspective taking, and reasoned that Level 2 perspective taking is an instantiation of explicit theory of mind.

One Context-Sensitive Perspective Taking System?

Recent research has questioned the distinctions between Level 1 and Level 2 visual perspective taking—and between implicit and explicit theory of mind more generally (see Westra, 2017; Carruthers, 2016, 2017; Christensen & Michael, 2015; Elekes, Varga, & Király, 2016, 2017; Surtees, Samson, & Apperly, 2016). Some researchers have challenged claims that Level 1 perspective taking is automatic and modular, as mounting evidence suggests it is influenced by participants' knowledge, goals, and emotional states (Carruthers, 2017; Westra, 2017; Todd & Simpson, 2016).[147] More relevant to this Chapter, however, is recent evidence that Level 2 perspective taking is much quicker and less effortful than is widely assumed.

Elekes, Varga, & Király (2016) adapted Surtees and colleagues' number perspective task such that it could be completed with another live, human participant. Participants were seated at a table that had a screen embedded in it. Participants in the "Individual" condition were the only ones sitting at the table, while participants in the "Joint" condition completed the study with a partner seated on the opposite side of the table. Each trial of the study proceeded as follows: A computerized female voice would say a number, such as "eight" or "nine." Then, the screen in the table would display a number, in one of several possible colors. Participants were told to complete one of two tasks. Participants completing the "perspective dependent" task pressed a key to indicate whether the number on the screen matched the number the voice said.

[147] Indeed, it has recently been suggested that *no* social cognitive process can be truly "automatic"—perhaps "spontaneous" is as close as it gets (Elekes, Varga, and Király, 2017).

Participants completing the "n-back" task, on the other hand, pressed a key to indicate whether the *color* of the stimulus matched the color of the previous stimulus. Importantly, participants in the Joint condition were always made aware of what task the other participant was doing; therefore, participants knew what aspects of the visual stimuli were relevant to their partners.

The critical results relate to the perspective dependent task. Elekes and colleagues found that participants in the Joint condition were slower to respond than participants in the Individual condition, but *only* on the trials for which (i) their partner was completing the perspective dependent task, and (ii) the number on the screen was one for which the partner's response would be different from their own (e.g. a 6 rather than an 8). In other words, Elekes and colleagues observed altercentric interference on the number perspective task: participants seemed to calculate their partner's Level 2 visual perspectives, even though it was not relevant to their individual goal, so long as they knew their partner was attending to the same features of the visual stimulus (i.e. the value of the numbers on the screen).

Surtees, Samson, & Apperly (2016) conducted a similar two-live-participants variation of the number perspective experiment, and got similar results. Notably, however, Surtees and colleagues employed a block-based turn-taking design in which Joint condition participants took turns being the "player" and the "observer." They found that altercentric interference did not emerge until after the first player's first block, which they took as confirmation that "it is not the biophysical features of another person that prompts us to take their perspective, but rather the context in which we act" (ibid. 50).

Most recently, Elekes, Varga, & Király (2017) applied their 2016 paradigm to groups of children at 8 years of age and 9.5 years of age. Previous developmental findings had indicated that children of these ages experience altercentric interference in Level 1 perspective taking (i.e.

the dot perspective task; Surtees & Apperly, 2012), but not in Level 2 perspective taking (i.e. the number perspective task; Surtees, Butterfill, & Apperly, 2012). But, when the Level 2 perspective taking task was modified to include a live human partner, children at both age 8 and age 9.5 exhibited altercentric interference in the same circumstances observed in Elekes, Varga, & Király (2016).

Collectively, the results of these studies "pose a considerable challenge for the classical two-system view that posits a strict dichotomy between automatically computed level-1 perspectives and offline, effortfully computed level 2 perspectives" (Elekes, Varga, & Király, 2017, p. 612). Rather, it seems that people can "compute both forms of visual perspectives in a quick but context-sensitive way, indicating that the two functions share more features than previously assumed" (ibid., 609). Both level 1 and level 2 perspective taking, it seems, can happen online, even in the absence of instruction, and can be fast and efficient, at least in certain contexts (e.g., given sufficient knowledge and motivation).

In light of these findings, one-system views of theory of mind are coming back into fashion (Carruthers, 2017; Westra, 2017; Elekes, Varga, & Király, 2017). Carruthers (2017) argues that a single theory of mind system can explain the findings to date more parsimoniously than a two-system account. He posits this one system operates "automatically" where it can, but depending on contextual factors (such as motivation) typically works together with domain-specific executive procedures (e.g. mental rotation) and domain-general resources (e.g. working memory).

Similarly, Westra (2017) suggests that, rather than two distinct mindreading systems, people are equipped with a general, efficient mindreading capacity that they deploy selectively, depending on context and goals. Recognizing that the mechanisms that allow people to

efficiently engage in Level 2 perspective are an open question, Westra posits strategies that exploit knowledge and/or long-term memory might allow people to circumvent the need to use working memory. Specifically, in the number perspective task, "once subjects learned that their partners' perspective systematically differed from their own (e.g. "If I see 6, he sees 9"), they would have been able to store that knowledge as a schema in long-term memory, where it would have been available for rapid retrieval" (p. 4572).

The emerging one-system accounts share a number of features, including their emphasis on the role context plays in determining perspective taking strategies and processes. Researchers are converging on the idea that perspective taking depends on presently-unknown contextual factors (Elekes, Varga, & Király, 2017). The experiments reported in this Chapter investigate what some of those contextual factors might be.

Spontaneous Perspective Taking in Single-Trial Studies

After a comprehensive review of the theory of mind literature, Westra (2017) concluded that "we only ever mindread as much as we have to" (p. 4577). Yet some findings are difficult to reconcile with this statement.

Most notably, Tversky and Hard (2009) conducted a single-trial experiment in which participants viewed one of three photographs printed on a piece of paper (see Figure 3.1). Participants in the "no person" condition saw a bottle and a book resting on top of a table in a room, with no person in the picture (Figure 3.1(c)). Participants in the "looking" condition saw a similar picture, except that there was a man on the opposite side of the table looking at the book (Figure 3.1(b)). Finally, participants in the "reaching" condition saw a picture similar to the "looking" picture, except the man on the opposite side of the table was extending his left arm

toward the book. Participants in all conditions wrote answers to a single question printed beneath the picture: "In relation to the bottle, where is the book?" A substantial minority of participants responded from the man's perspective: 29% in the reaching condition and 22% in the looking condition, versus only 3% in the no person condition. In a second experiment, all participants viewed the reaching photograph (Figure 3.1(a)), but were asked one of four different questions, two of which mentioned action by the man in the photograph ("In relation to the bottle, where does he place the book?" and "In relation to the bottle, where is the book placed?") and two of which did not ("In relation to the bottle, where is his book?" and "In relation to the bottle, where is the book?"). Fifty percent of participants responding to action questions answered from the man's perspective, versus 33% of participants responding to non-action questions. Tversky and Hard interpreted their findings as a challenge to the perceived primacy of the egocentric perspective, concluding that it is often natural to spontaneously adopt another's perspective.

Figure 3.1. Figure 1 from Tversky & Hard (2009) depicting conditions in that study.

Todd et al. (2015) investigated the effects of incidental emotion—specifically anxiety—on visual perspective taking (among other theory-of-mind related tasks). Using the looking stimulus from Tversky and Hard (2009) and a similar one-trial design, it appears that Todd et al. (2015) found comparable levels of visual perspective taking. However, Todd et al. (2015)

reported only responses from participants' own egocentric perspectives (72.3% in the anxiety condition, 50.0% in the anger condition, and 45.5% in the neutral condition)—it is not clear from the manuscript whether all of the other participants responded from the perspective of the man in the photograph, or whether some subset of participants responded using neither perspective. In any event, Tversky and Hard (2009) and likely Todd et al. (2015) found that a substantial share of participants engage in level 2 perspective taking, even in situations where they certainly do not have to do so.

One important commonality between Tversky and Hard (2009) and Todd et al. (2015) is their use of single-trial designs. Most of the level 2 perspective taking research to date has used repeated-trial designs. While repeated-trial designs offer some advantages, they also have significant drawbacks when attempting to study "spontaneous" phenomena. First, participants might detect and respond to task demands that become apparent across multiple trials. For example, as noted above, Surtees, Samson, & Apperly (2016) reported what they characterized as "spontaneous" level 2 perspective taking, but this "spontaneous" perspective taking did not occur until participants gained familiarity with the task through the first block of trials. Is perspective taking in such circumstances "spontaneous" or an artifact of a particular research design? A single-trial design allows researchers to largely avoid this complicated issue; contextual factors that matter in a single-trial design, by definition, matter in a participant's first (and only) exposure. Second, given repeated exposures to the same perspective taking ambiguity (e.g. 6 vs. 9), participants might effectively automatize the information needed to calculate other perspective so that doing so becomes less effortful (Westra, 2017). With a single-trial design, participants do not have the opportunity to outsource effortful perspective taking to long-term memory. Thus, single-trial designs are in many respects better suited for investigating whether

perspective taking occurs spontaneously, and the evidence from single-trial studies suggests it does.

The Present Experiments

Prior research demonstrates that a substantial portion of people will spontaneously adopt another's level 2 visual perspective, even when it is not necessary to do so. In this Chapter, I present five experiments that build on these findings by exploring factors that facilitate, or suppress, such spontaneous level 2 perspective taking. I focus specifically on the roles of *context* and *capacity*.

With respect to context, prior work reveals three noteworthy things. First, Tversky and Hard (2009) observed that calling attention to another person's *action* (based either on visual cues or question framing) facilitates spontaneous level 2 perspective taking. Second, findings from Elekes, Varga, and Király (2016, 2017) suggest that the presence of a *live person*, rather than an avatar, facilitates spontaneous level 2 perspective taking. Third, Elekes, Varga, and Király (2016, 2017) also observed that knowledge of the other person's *goals* affects spontaneous perspective taking. Collectively, these findings suggest a broader insight: people's proclivity for spontaneous perspective taking might depend on their perceptions of, or attention to, the relevant other's *agency*. That is, the more deeply one conceptualizes the other person or entity as an agent with the capacity for goal-directed action (see Gray, Gray, & Wegner, 2007; Johnson, 2003), the more likely one is to spontaneously adopt that person or agent's level 2 perspective. To explore this possibility, my experiments assess whether participants spontaneously adopt the perspective of a schematic "happy face"—a stimulus that participants could conceivably conceptualize as a representation of an agent, or as a collection of simple

shapes (see Figure 3.2). My first two experiments investigate how manipulating the salience of the happy face's agency affects perspective taking.

With respect to capacity, prior work has explored the relation between cognitive load and level 1 perspective taking (Qureshi, Apperly, & Samson, 2010), and between cognitive load and implicit theory of mind more generally (e.g. implicit false-belief tracking in Schneider, Lam, Bayliss, & Dux, 2012; disambiguating referential communication in Cane, Ferguson, & Apperly, 2017). But the role of load is underexplored with respect to Level 2 perspective taking.[148] After finding evidence of spontaneous level 2 perspective taking, Elekes, Varga, and Király (2016) called for future research to test whether cognitive load interferes with such perspective taking, thus probing the efficiency of the process. To my knowledge, my experiments 3 through 5 are the first to provide these tests, exploring the effects of both visual and auditory working memory load.

Each experiment featured a single-trial design, similar to Tversky and Hard (2009). In each, participants viewed an image of a schematic "happy face" on a computer screen (see Figure 3.2),[149] then responded to a critical question about the relative position of a red square in the image by typing an open-ended response. Responses were coded to reflect whether participants answered from their own spatial perspective or from the (opposite) perspective of the happy face.[150] From study to study, we varied the instructions and preliminary tasks that

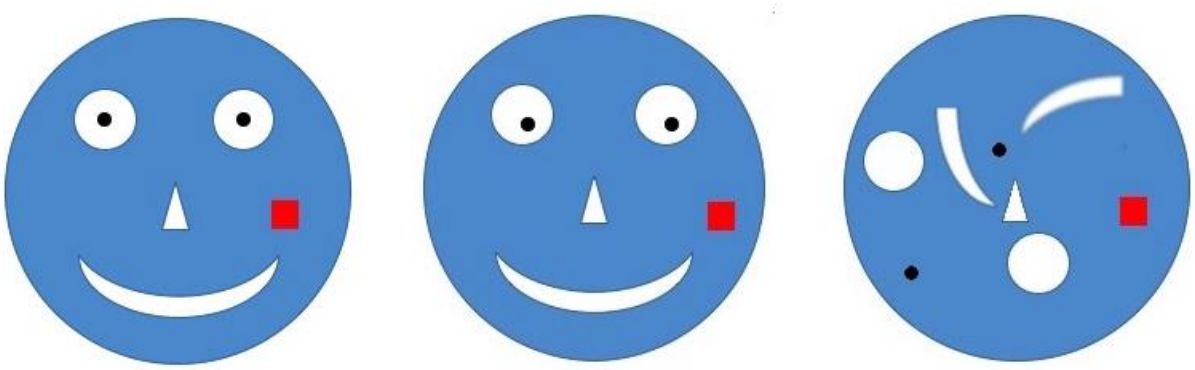
[148] This is likely because, until recently, it has been largely viewed as slow, resource-intensive process that would surely be suppressed with load.
[149] Some participants in the first experiment viewed a control, scrambled version of the happy face image rather than a recognizable happy face.
[150] A small number of responses did not indicate any perspective and were excluded from analyses.

preceded the critical question to investigate how context and capacity influenced spontaneous perspective taking.

Experiment 1

As discussed above, prior research suggests that whether people spontaneously adopt others' perspectives depends, in part, on perceptions of the others' agency. Specifically, I hypothesize that people are more likely to adopt another's visual perspectives when focused on the other's agency—and, therefore, that spontaneous level 2 perspective taking will increase with the salience of the other's agency.

Experiment 1 provides a preliminary test of this hypothesis. Prior research on level 2 perspective taking has used photographic stimuli or quasi-realistic avatars displayed on computer screens. Here, however, I probed level 2 perspective taking using a simple schematic happy face comprised of eight simple shapes (see Figure 3.2). This afforded the opportunity to nudge participants toward conceptualizing the stimulus as either an agent or as a collection of shapes—both through the visual arrangement of the stimulus and through the framing of questions about it.



Standard Face Looking Face Control

Figure 3.2. Examples of the "standard face," "looking face," and control stimuli. The locations of the red square (on the left or right cheek) were counterbalanced.

The use of the schematic happy face stimulus also offered another advantage. The eight shapes could be rearranged in a manner that did not resemble a face (see Figure 3.2, control stimulus). Thus, the stimulus enabled me to control for the perceptual features of the stimulus while manipulating whether it depicted an agent.

Method

Participants. 110 participants completed the experiment through Amazon Mechanical Turk. Due to the possibility that some participants may attempt to complete the study more than once, an *a priori* exclusion rule was established: participants would be excluded if (i) they had the same Mechanical Turk ID as another participant, (ii) their response was associated with the same IP address as another participant, or (iii) they responded from the same geographic coordinates as another participant. Pursuant to this rule, 4 responses were excluded.

The final sample of 106 participants included 53 men and 53 women, ranging in age from 21 years to 68 years with an average age of 35.39 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedure. After consenting to the study, participants read a brief set of general instructions. The instructions informed participants that they would see an image and be asked a question about it, and told participants to answer the question in their own words.

After reading these instructions, participants were shown one of three images: a standard happy face with a red square on one of its cheeks; a "looking" happy face, with eyes gazing toward the red square on one cheek; or a scrambled-face control (see Figure 3.2). The location of the red square (either on the left cheek or right cheek) was counterbalanced. One of two questions appeared beneath the image: "relative to the triangle, where is the red square?" (the "non-agency prompt") or "relative to the nose, where is the red square?" (the "agency prompt") (see Tversky & Hard, 2009).

Note that, because participants in the scrambled-face control condition were unlikely to recognize the triangle as a "nose," they were always given the "non-agency prompt." Thus, participants could be assigned to one of five conditions: standard face with non-agency prompt (N = 20); standard face with agency prompt (N = 22); looking face with non-agency prompt (N = 21); looking face with agency prompt (N = 19); and scramble face control with non-agency prompt (N = 22).

Participants typed the response to the critical question into an open-ended text box. After submitting their responses, participants were asked a series of demographic questions about themselves.

Results and Discussion

Two responses were excluded for failure to answer the question in a way that indicated perspective (e.g. "on the face") (see Tversky & Hard, 2009), leaving 104 responses for analysis.

Condition had a significant effect on level 2 perspective taking, $X^2(4, 104) = 14.51$, $p = .006$. This effect was driven by prompt formulation rather than differences in the image. Specifically, examining the four non-control conditions in which participants saw a face, participants were more likely to take the face's perspective in response to the agency prompt (36.59%) than to the non-agency prompt (9.75%), $X^2(1, 82) = 8.289$, $p = .004$.

The image manipulation (control vs. standard face vs. looking face) had no effect on perspective taking. The lack of an effect may have been due to the lack of limbs in the happy face image (cf. Tversky & Hard, 2009). Although the looking face is gazing at the red square, there is no plausible way for the face to act on the square, even within the context of the picture. Looking alone may have been too subtle to suggest the face has interests or goals related to the square, or otherwise draw attention to the face as an agent.

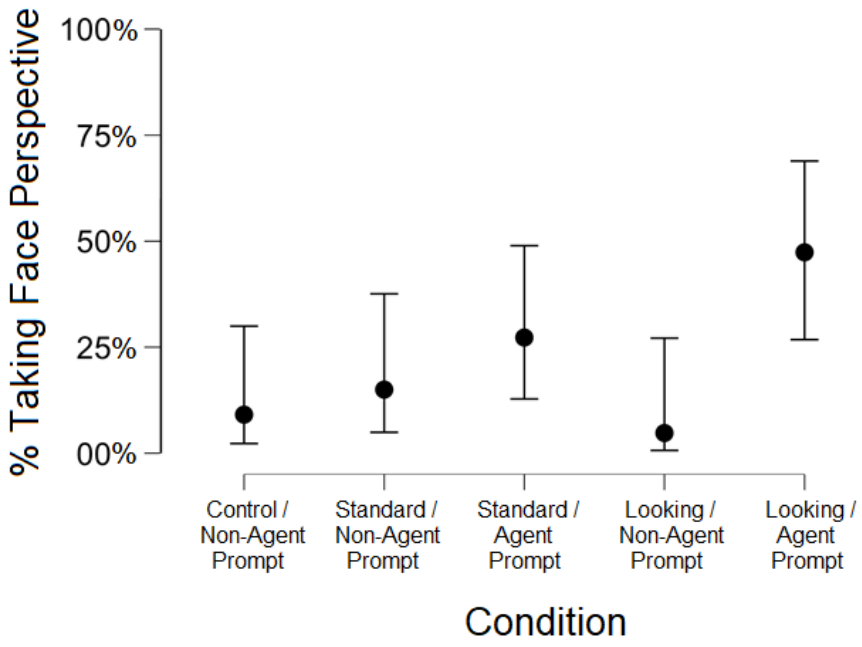


Figure 3.3. Percentage of participants that engaged in spontaneous level 2 perspective taking by condition in Experiment One. Error bars reflect 95% confidence intervals.

As expected, the highest level of perspective taking was found when participants viewed the looking face and responded to the agency prompt. Participants in this condition responded from the face's perspective 47% of the time (see Figure 3.3). This rate of perspective taking was similar to that observed by Tversky & Hard (2009) using photographic stimuli with human agents (between 22% and 50%, depending on condition). This is evidence that participants will engage in spontaneous level 2 perspective taking for even abstracted agentive stimuli. The combination of the looking face stimulus and the agency prompt is used as the stimulus in the remainder of the experiments.

Experiment 2

The success of the prompt manipulation in Experiment 1 suggested that participants' consideration of the happy face's agency facilitates spontaneous perspective taking. Experiment 2 built on that finding. Specifically, in Experiment 2, participants completed one of three preliminary tasks before answering the critical question about the location of the red square. These preliminary tasks were designed to invite consideration of the happy face stimulus either as an agent or as a collection of shapes.

Method

Participants. 150 participants completed the experiment through Amazon Mechanical Turk. Four were excluded pursuant to the exclusionary rule described in Experiment 1.

The final sample of 146 participants included 50 men, 93 women, and 3 participants who preferred not to say, and ranged in age from 20 years to 74 years with an average age of 35.53 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedures. Stimuli and procedures resembled Experiment 1 except that participants were assigned to answer one of three preliminary questions before answering the critical question.

Participants in the "Agency Framing" condition were asked what emotion the face was experiencing. All participants responded with a variant of "happiness." Participants in the "Shape Count Framing" condition were asked to count the number of white shapes in the image. Most participants in this condition responded with the correct count of four. Finally, participants in the "Face Distance Framing" condition were asked a question about distances between

components of the face: “Which is greater, the distance from the nose to the mouth or the distance from the nose to the red square?”

Results and Discussion

Sixteen responses were excluded for failure to answer the question in a way that indicated perspective (e.g. “on the face”), leaving 130 participants for analysis.

As in Experiment 1, perspective taking varied significantly across conditions, $X^2(2, 130) = 7.51$, $p = .023$. Specifically, there was a significant difference between the Agency Framing condition, in which 49% of participants (19 of 39) took the face’s perspective, and the Shape Count Framing condition, in which only 20% of participants (9 of 44) took the face’s perspective, $X^2(2, 83) = 7.39$, $p = .007$. The Face Distance Framing (15 of 47) condition did not differ significantly from either of the other two conditions (see Figure 3.4).

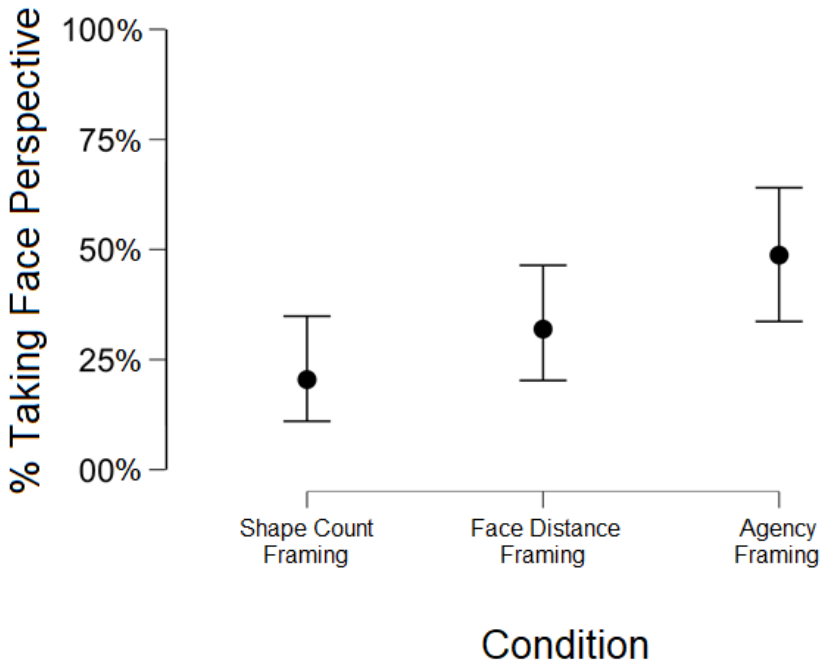


Figure 3.4. Percentage of participants that engaged in spontaneous level 2 perspective taking by condition in Experiment Two. Error bars reflect 95% confidence intervals.

Experiment 2 provides additional evidence that cues designed to prompt participants to consider the happy face's agency tend to facilitate spontaneous perspective taking, whereas cues designed to prompt participants to focus on non-agentive features tend to suppress it. There is one noteworthy alternative interpretation of these findings: the difference between the "Agency Framing" condition and the "Shape Count Framing" condition may have been less about consideration of the face's agency than about whether the face image was processed holistically. The question about the happy face's mood may have invited participants to process the happy face as an integrated whole, whereas the shape question may have invited participants to process the image as a collection of shapes. That said, the Face Distance Framing condition arguably invites holistic processing as much or more than the Agency Framing condition, as answering required participants to attend to the relation between two aspects of the face (whereas the Agency Framing condition arguably just drew attention to one feature—the smile). In any event, future research could disentangle these possibilities by contrasting the Agency Framing with a perceptual framing question asked about the face as a whole, such as "what percentage of the computer screen does the image above occupy?"

Experiment 3

Experiments 1 and 2 examined the role that context (and, specifically, the salience of agency) plays in spontaneous level 2 perspective taking. Experiments 3 through 5 focus on the role of capacity.

If spontaneous level 2 perspective taking draws on domain-general cognitive processes, then placing participants under cognitive load should suppress it by depleting those resources (Carruthers, 2015). Research in other theory-of-mind domains suggests that load often impairs

theory of mind processes (e.g. Lin, Keysar, & Epley, 2010)—both those involving explicit, "full-blown" theory of mind (cf. Qureshi, Apperly, & Samson, 2010) and, in some instances, those involving "minimal" or "implicit" theory of mind (Schneider, Lam, Bayliss, & Dux, 2012). There is some reason to believe load will suppress level 2 visual perspective taking, in particular. Specifically, to the extent that anxiety operates as a drain on domain-general cognitive resources akin to load (see Vytal, Cornwell, Arkin, & Grillon, 2012), Todd et al.'s (2015) finding that anxiety suppresses level 2 perspective taking, even in a single-trial study, leads one to expect load to do the same.

In Experiment 3, participants were placed under visual working memory load using a matrix task before viewing and responding to the critical location question. I hypothesized that this load would suppress perspective taking (see Carruthers, 2016, 2017).

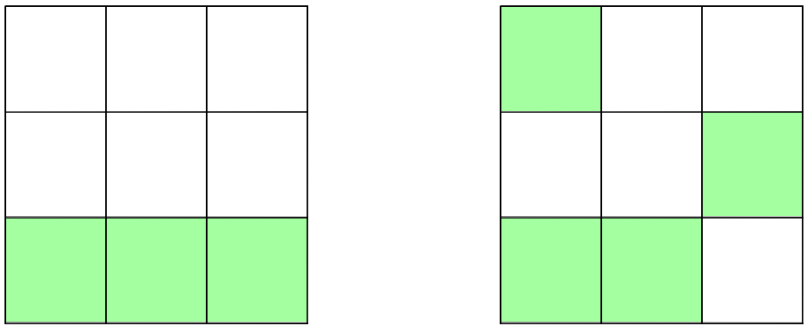
Method

Participants. 122 participants completed the experiment through Amazon Mechanical Turk. One response was excluded pursuant to the rules described above in Experiment 1.

The final sample of 121 participants included 46 men, 74 women, and 1 participant who preferred not to say, and ranged in age from 18 years to 69 years with an average age of 38.44 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedures. Stimuli and procedures resembled Experiment 1 except that participants encoded a visual working memory load before they viewed the face image and answered the critical location question. For the load task, participants were told that a 2-color matrix pattern would appear on their screen for 2 seconds, then disappear. Participants were

instructed to memorize the matrix pattern so that they could re-create it later in the experiment (see Maldonado, 2016). The matrix pattern the participants saw varied based on whether participants were assigned to the "Low Load" or "High Load" condition (ibid.; see Figure 3.5). In the "Low Load" condition, three squares of the matrix were shaded in a straight line. In the "High Load" condition, four squares were shaded in various positions around the matrix.



Low Load High Load

Figure 3.5. Low load and high load matrices for visual working memory load task.

After the matrix pattern was displayed for 2 seconds, it disappeared from the screen and was replaced with the face image and were given the agency prompt ("relative to the nose, where is the red square?"). After submitting their response to the question, participants re-created the matrix by clicking on the squares to shade them in a graphical user interface. Out of abundance of caution, participants who failed to correctly re-create the matrix were excluded from analysis (as it is difficult to ensure they were experiencing load since they did not successfully deploy the retained information) (see Bago & De Neys, 2019).

Results and Discussion

Three responses were excluded for failure to answer the question in a way that indicated perspective (e.g. "on the face"). Of the remaining 117 participants, 15 were excluded for failing the re-create the matrix they viewed during the load task: 3 of 60 participants were excluded for failing to re-create the Low Load matrix, while 12 of 57 participants were excluded for failing to re-create the High Load matrix. This left a total of 102 participants for analysis: 57 in the Low Load condition and 45 in the High Load condition. Perspective taking did not vary across load conditions (Low Load: 13 of 57, 22.8%; High Load: 8 of 45, 17.8%), $X^2(1, 102) = .389$, $p = .533$.[151]

Experiment 4

Experiment 4 resembled Experiment 3 but investigated the role of *auditory* working memory load rather than visual working memory load. Auditory (or verbal) load is widely thought to draw upon a different store of working memory than visual load (see Baddeley, Grant, Wight & Thomson, 1975) and its effects on tasks are often considered separately (e.g. Woodman, Vogel, & Luck, 2001). There is reason to suspect that auditory load will matter in my perspective taking paradigm, as the phenomena being measured are "high-level and cognitive, reflected in and affected by language" (Tversky & Hard, 2009, p. 128). In addition to switching the type of load, Experiment 4 added a "No Load" condition, to account for the possibility that any degree of load suppresses perspective taking.

[151] The difference remains near-significant if the participants who failed the load task are included (13 of 60 = 21.7% in low load versus 11 of 58 = 19.0%, $X^2(1, 118) = .133$, $p = .716$).

Method

Participants. 149 participants completed the experiment through Amazon Mechanical Turk. Five responses were excluded pursuant to the rules described above in Experiment 1.

The final sample of 144 participants included 73 men, 69 women, and 2 participants who preferred not to say, and ranged in age from 20 years to 66 years with an average age of 34.85 years. All of the participants were U.S. citizens and native English speakers. Participants were compensated for their time.

Stimuli and Procedures. Stimuli and procedures resembled Experiment 3 with two exceptions. First, some participants in this experiment were assigned to a "No Load" control condition. In this condition, participants proceeded directly from the initial general instructions to the looking face image and agency framing question, with no load task. This condition was identical to the "Looking Face with Agency Prompt" condition in Experiment 1.

Second, in Experiment 4, the participants in the load conditions completed an auditory working memory load task instead of a visual working memory load task. Specifically, after completing a speaker test to confirm their audio was working, participants were instructed to listen to a series of digits, which would play over their speakers. Participants were told to memorize the series of digits so that they could enter it later in the experiment. The series of digits the participants heard varied between a "Low Load" condition (2 digits) and a "High Load" condition (6 digits). After the series of digits played, the looking face image and agency prompt appeared on the screen.

Once participants submitted their response to the agency prompt, they were prompted to type the series of digits they heard in the auditory working memory load task. Similar to

Experiment 3, participants who failed to correctly type the series of digits were excluded from analysis out of an abundance of caution (see Bago & De Neys, 2019).

Results and Discussion

Of the 144 participants in the sample, 36 were excluded for failing the re-create the digit span they heard during the load task: 11 of 53 participants in the Low Load condition and 25 of 48 participants in the High Load condition. This left a total of 108 participants for analysis: 42 in the Low Load condition, 23 in the High Load condition, and 43 in the No Load condition.

Perspective taking in Experiment 4 did not vary across load conditions, $X^2(2, 108) = 1.830$, $p = .401$. Participants were descriptively most likely to engage in spontaneous level 2 perspective taking in the No Load condition, as predicted (15 out of 43 participants, or 35%), but that did not differ significantly from the Low Load (10 out of 42 participants, or 24%) or High Load (5 out of 23 participants, or 22%) conditions (see Figure 3.6).[152]

[152] The results are essentially the same if participants who failed the load task are included (No Load: 15 of 43, 35%; Low Load: 14 of 53, 26%; High Load: 10 of 48, 21%; $X^2(2, 144)=2.286$, $p=.319$).

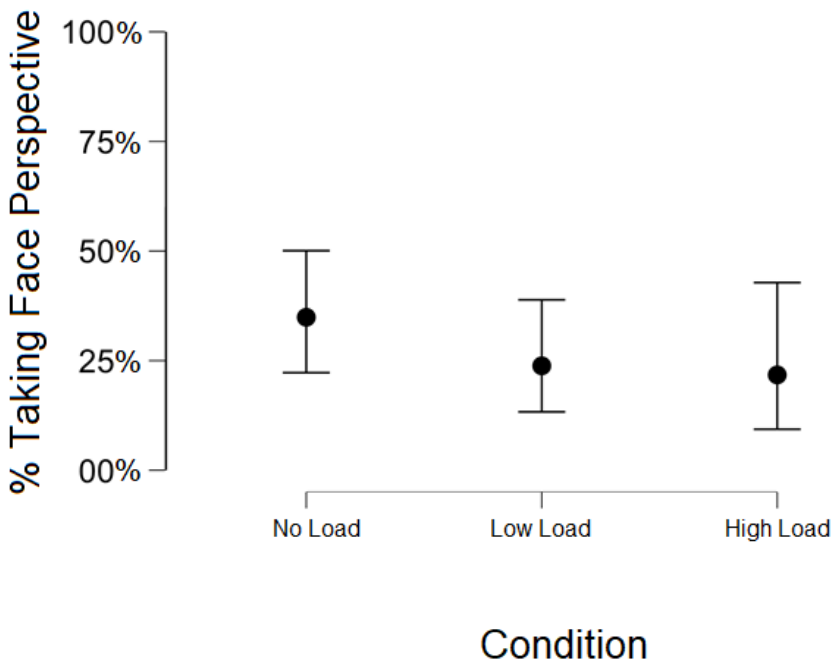


Figure 3.6. Percentage of participants that engaged in spontaneous level 2 perspective taking by condition in Experiment Four. Error bars reflect 95% confidence intervals.

I conducted a post hoc analysis combining the Low Load and High Load conditions, allowing me to compare participants under (any amount of) load to No Load participants. The results were not statistically significant, $X^2(1, 108) = 1.798$, $p = .180$.

However, given the pattern of results in Experiment 4, it is plausible that there is a latent effect of auditory load that simply did not reach significance in the sample. The No Load produced, descriptively, the highest percentage of spontaneous perspective taking, 14% higher than the percentages observed in either the Low Load or High Load condition. Thus, Experiment 5 was designed to further probe the influence of load on perspective taking.

Experiment 5

While I did not examine any effects of cognitive load in Experiments 3 or 4, an examination of absolute levels of perspective taking raises the possibility load might matter. Across the previous experiments, 24 of 62 participants in the No Load condition (38.7%) engaged in spontaneous perspective taking, while participants under any level of visual or auditory load collectively engaged in perspective taking 21.5% of the time. Thus, I designed Experiment 5 to further investigate the effect of both auditory load and visual load in a larger sample.

Method

Participants. 352 participants[153] completed the experiment through Amazon Mechanical Turk. 27 responses were excluded pursuant to the rules described above in Experiment 1 (nine due to repeated IP addresses and 18 due to repeated geographic locations).

The final sample of 325 participants included 156 men, 167 women, and 2 participants who preferred not to say, and ranged in age from 18 years to 88 years with an average age of 35.70 years. All of the participants were U.S. citizens. All reported speaking English and 310 were native English speakers. Participants were compensated for their time.

Stimuli and Procedures. Experiment 5 featured 5 conditions: low visual load, high visual load, low auditory load, high auditory load, and no load. Participants in visual load conditions experienced the same stimuli and procedures as participants in the corresponding conditions in

[153] Calculations with G*Power indicate that using a chi-square analysis with 4 degrees of freedom, to detect an effect of size $w = .3$ at $\alpha = .05$ and power = .95, a total sample size of 207 is needed. More participants were recruited due to the high exclusion rates seen in Experiments 3 and 4. After exclusions, 245 participants were analyzed in Experiment 5.

Experiment 3. Participants in the auditory load conditions and the no load condition experienced the same stimuli and procedures as participants in the corresponding conditions in Experiment 4.

Results and Discussion

As an initial step, 32 responses were excluded for failure to answer the question in a way that indicated perspective (e.g. "on the face"). The remaining 293 participants were distributed across conditions as follows: 59 in the low visual load condition, 59 in the high visual load condition, 62 in the low auditory load condition, 55 in the high auditory load condition, and 58 in the no load condition.

Subsequently, participants who failed to re-create the matrix they saw (in the visual load conditions) or digit span they heard (in the auditory load conditions) during the load task were also identified and excluded. As expected, the high load conditions proved more difficult than the low load conditions. Only two of 62 participants failed to re-create the two-digit digit span in the low auditory load condition, versus 32 of 55 participants who failed to re-create the six-digit span in the high auditory load condition. Similarly, no participants failed to re-create the low-load matrix, whereas 14 of 59 failed in the high visual load condition.

Among the final sample of 245 participants, perspective taking did not vary significantly across conditions, $X^2(4, 245) = 5.634$, $p = .228$. As expected, and as in Experiment 4, participants were descriptively most likely to engage in spontaneous level 2 perspective taking in the No Load condition (17 out of 58 participants, or 29.3%). But the No Load condition did not differ significantly from the Low Visual Load (11 out of 59 participants, or 18.6%), the High

Visual Load condition (9 of 45, or 20.0%), the Low Auditory Load condition (8 of 60, or 13.3%), or the High Auditory Load condition (3 of 23, or 13.0%) (see Figure 3.7).[154]

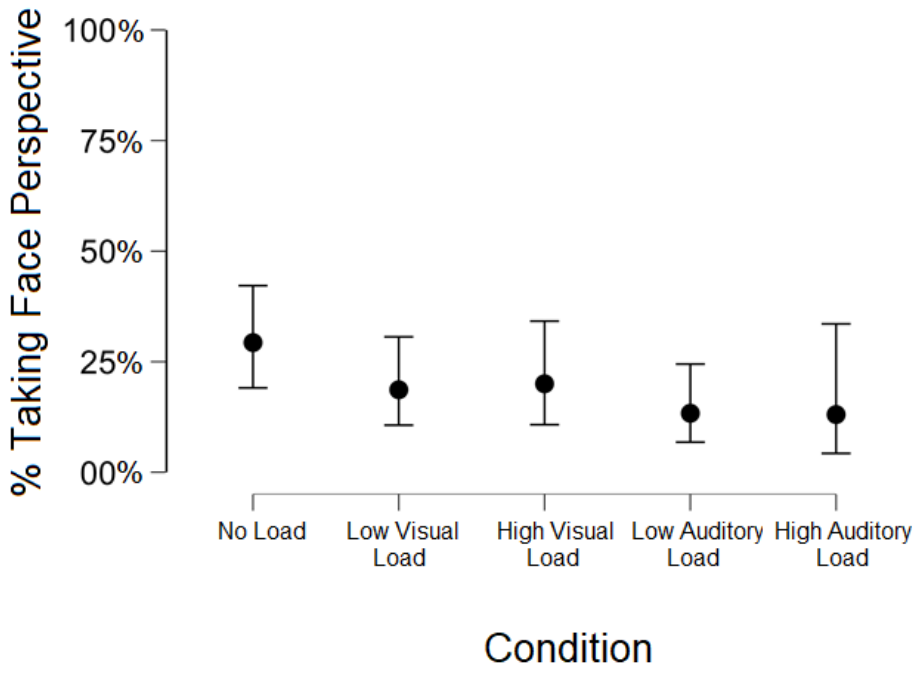


Figure 3.7. Percentage of participants that engaged in spontaneous level 2 perspective taking by condition in Experiment Five. Error bars reflect 95% confidence intervals.

Following this null result, I conducted a post hoc analysis combining participants across all four load conditions and comparing them to participants in the No Load condition. This revealed that participants in the No Load condition were more likely to perspective take than participants under any amount of load, $X^2(1, 244) = 4.473$, $p = .034$.[155] Additional post hoc

[154] The results are the same if the participants who failed the load tasks are included in the analysis (No Load = 17 of 58, 29.3%; Low Visual Load = 11 of 59, 18.6%; High Visual Load = 12 of 59, 20.3%; Low Auditory Load = 9 of 62, 14.5%; High Auditory Load = 12 of 55 = 21.8%; $X^2(4, 293) = 4.242$, $p = .374$).

[155] The difference falls just short of statistical significance if participants who failed the load tasks are included in the analysis, $X^2(1, 293) = 3.163$, $p = .075$).

analyses explored whether individual differences might explain some portion of the variability in participants' spontaneous level 2 perspective taking. However, a series of logistic regressions revealed no significant relations between perspective taking and age, gender, education level, or parents' education level.

General Discussion

The studies reported in this Chapter reveal three important things. First, a substantial share of participants spontaneously engage in Level 2 perspective taking for a schematic happy face stimulus comprised of simple shapes. Across all studies, participants under no cognitive load answered the critical perspective-taking question (about the position of a red square relative to the nose of a schematic face) from the face's perspective 34.2% of the time, even though their task did not require them to account for the face's perspective. This rate of perspective taking is comparable to the rate that Tversky and Hard (2009) observed using photographs of humans (which ranged from 22% to 50%, depending on question framing), suggesting such spontaneous perspective taking occurs for a range of abstracted agentive stimuli.

Second, my studies demonstrate that prompting participants to consider a stimulus's agency facilitates spontaneous perspective taking. In Experiment 1, participants were more likely to respond from the schematic face's perspective when the critical question was framed in agentive language (e.g. referring to the face's "nose") than when it was framed in non-agentive language (e.g. referring to the "triangle"). Experiment 2 indicated that preliminary tasks that invite participants to think of a stimulus as an agent can also have facilitative effect. Participants asked to consider the face's mood (versus count the number of shapes in the schematic face) were more likely to perspective take when they subsequently read and responded to the critical question.

Third, though evidence was mixed, my studies collectively provide some support for the proposition that cognitive load suppresses spontaneous Level 2 perspective taking. Evidence is limited at the level of individual experiments. Experiments 3 and 4 showed no effect of auditory or visual load on perspective taking. In Experiment 5, there were no significant differences in perspective taking across no load, visual load, and auditory load conditions, though combining load conditions revealed that participants under no load were more likely to perspective take than participants under (any) load. However, the design of the studies allows for pooling of data: the low and high visual load conditions in Experiment 3 were identical to those in Experiment 5; the low and high auditory load conditions in Experiment 4 were identical to those in Experiment 5; the no load conditions in Experiments 4 and 5 were identical to the "looking / agent prompt" condition in Experiment 1.

When data is pooled across these studies, some notable patterns emerge. Specifically, across studies, participants in no load conditions were more likely to perspective take (41 of 120, 34.2%) than participants under any load (67 of 354, 18.9%), $X^2(1, 474) = 11.831$, $p < .001$). Indeed, participants under no load exhibited more spontaneous perspective taking than participants in each of the other four load conditions, responding from the face's perspective more than participants under low auditory load (18 of 102, 17.6%; $X^2(1, 222) = 7.71$, $p = .005$, $w = .186$), high auditory load (8 of 46, 17.4%; $X^2(1, 166) = 4.499$, $p = .034$, $w = .165$), low visual load (24 of 116, 20.7%, $X^2(1, 236) = 5.368$, $p = .021$, $w = .389$), or high visual load (17 of 90, 18.9%, $X^2(1, 210) = 6.005$, $p = .014$, $w = .169$). The pooled data thus suggests a difference between no load and the load conditions, though the effect sizes are rather modest.

My findings that agency prompts and perhaps load affect the prevalence of level 2 perspective taking has implications for the one-versus-two system debate in the perspective

taking literature (and the theory of mind literature more broadly). The conventional two-system view is that level 2 perspective taking is an instantiation of the explicit theory of mind system, and as such, an effortful process—one people should not engage in without a reason. Yet my findings, together with others (Tversky & Hard, 2009; Elekes, Varga, and Király, 2016, 2017) demonstrate that people will often engage in level 2 perspective taking even where there is no reason to expend the cognitive effort. This sort of "spontaneous" perspective taking has traditionally been viewed as a hallmark of implicit (and perhaps automatic) processes. But my findings are also problematic for the idea that level 2 perspective is implicit or automatic.

Conventionally, implicit or automatic processes are thought to be unaffected by information from other systems; the stimulus itself should trigger the process (Apperly & Butterfill, 2009). Yet in my studies, whether participants responded from the schematic face's perspective depended on more than the mere presence of the schematic face. Rather, perspective taking was affected by cues that prompted consideration of the schematic face as an agent, and (to a degree) by load on domain-general cognitive resources. A critic might argue that it is possible my manipulations affected participants' responses without affecting the underlying perspective taking processes. That is, participants may have mentally computed both their own perspective and the face's perspective, but *selected* which perspective they attended to or which answer they selected differently based on differences between conditions (cf. Qureshi, Apperly, & Samson, 2010). However, I am not aware of any theoretical bases that would suggest this possibility; once the cognitive effort to compute competing perspectives has been expended, it is unclear what would then push participants to select one possible perspective over the other in the absence of any communicative purpose.

Ultimately, my findings are most consistent with the (re-)emerging idea of a single, flexible perspective-taking system (Westra, 2017; Elekes, Varga, & Király, 2017). They suggest that, rather than fitting neatly into the category of an implicit or explicit process, level 2 visual perspective taking lies somewhere in between—not modular, yet frequently deployed online, without any instruction or need. The last point bears emphasis, as it contradicts the idea that "we only ever mindread as much as we have to" (Westra, 2017, p. 4577). Rather, my findings suggest that the often-useful social tendency to account for others' perspective is an *overlearned* process drawing on domain-general resources.

Overlearned social tendencies are so ingrained that people deploy them "mindlessly," launching the script without deeper consideration of context. Nass and Moon (2000) used the idea to explain why people apply social norms such as politeness and reciprocity in interactions with computers. Nass, Moon, and Carney (1999) found that participants responded to questions about Computer A's performance more positively when the questions were asked by Computer A than when they were asked by Computer B or in a paper-and-pencil format, despite their uniform rejection of the idea that computers have feelings that necessitate polite responses. They reasoned that their participants were taking a human social convention—positively biasing face-to-face evaluations of other people's performance—and mindlessly applying it in a superficially-similar situation involving a computer as a stand-in for the second person. Similarly, in my studies, it seems people are taking a human social convention—relaying directions using others' perspectives (Mainwaring, Tversky, Ohgishi, & Schiano, 2003)—and mindlessly applying it in a superficially-similar situation involving a schematic face as a stand-in for the second person.

Appendix A. Appendix to Chapter 1.

Vignettes Used in Experiments One-Two

Windsor v. International Computers
The plaintiff, Janet Windsor, is suing the defendant, International Computers, claiming that International Computers' negligence caused her injury.

Janet Windsor developed a rare form of skin cancer. After a long course of painful chemotherapy, doctors were able to cure the cancer.

Janet, who has worked as a secretary for years, believed that her cancer had been caused by the computer monitor that she used at her job. That monitor was manufactured by International Computers.

International Computers is a company that manufactures components of computer systems, including monitors. The type of International Computers monitor that Janet Windsor used, the IC5000, emits small amounts of a certain type of radiation that most doctors believe can cause skin cancer.

International Computers was aware that the IC5000 monitors emitted this radiation before putting the monitors on the market. International Computers was also aware that most doctors believed the radiation can cause skin cancer.

International Computers could have used an enhanced manufacturing technique to reduce the IC5000's radiation emissions before putting the monitors on the market. However, after reviewing the relevant manufacturing and medical information and meeting with safety and health consultants, the company chose to continue with its regular manufacturing technique.

Assume the following facts are true:

Given all of the information available at the time, 90% of companies in International Computers' position would have used an enhanced manufacturing technique to reduce radiation emissions before putting their monitors on the market.

International Computers' choice to use the regular manufacturing technique rather than the enhanced manufacturing technique led to an expected increase of three (3) cases of skin cancer among monitor users, each of which would be expected to cause $100,000 of damage.

This means that, if it had used the enhanced manufacturing technique, International Computers would have been expected to save $300,000 of costs to other members of society. Using the enhanced manufacturing technique would have cost International Computers $150,000 more than using the regular technique.

Vaughan v. Menlove Farms

The plaintiff, Vincent Vaughan, is suing the defendant, Menlove Farms, claiming that Menlove Farms' negligence caused him injury.

Vincent Vaughan works as a farmer in the western United States. He has a profitable farm, growing and selling corn, barley, wheat, and hay, as well as livestock.

Vincent is a careful farmer, and one thing he is very concerned about is keeping his hay dry. Moist hay is more likely to catch fire, and hayfires are a big risk to his business.

Vincent bought a special piece of hay-drying farm equipment. Vincent uses the equipment to make sure his hay is dry before storing it, reducing the risk of hayfires on his farm.

Menlove Farms operates a profitable farming business of its own on the property right next to Vincent Vaughan's farm.

Menlove Farms stores its hay in a large barn near the border between its property and Vincent's property.

Menlove Farms does not use special farm equipment to help ensure its hay is dry. Menlove Farms considered buying hay-drying farm equipment about a year ago, but after reviewing the relevant information and meeting with consultants, the company decided not to buy the equipment.

A few months ago, some moist hay in the Menlove Farms barn caught fire. The fire spread from the Menlove Farms barn to Vincent's property, badly damaging Vincent's crops and destroying several of Vincent's barns.

Assume the following facts are true:

Menlove Farms' choice not to buy special hay-drying farm equipment led to a to 1% greater risk of a hayfire. Such a hayfire would be expected to cause $5,000,000 of damage.

This means that, if it had purchased the special hay-drying farm equipment, Menlove Farms would have been expected to save $50,000 of costs to other members of society. Purchasing the special hay-drying farm equipment would have cost Menlove Farms $75,000.

Given all of the information available at the time, 10% of companies in Menlove Farms' position would have purchased the special hay-drying farm equipment.

Pendleton v. Dolman Transportation

The plaintiff, Patrick Pendleton, is suing the defendant, Dolman Transportation, claiming that Dolman Transportation's negligence caused him injury.

Dolman Transportation is a trucking company. About a year ago, Dolman Transportation expanded into the long-distance chemical hauling business.

Dolman Transportation bought several special chemical-hauling trucks for its fleet. When choosing its chemical-hauling trucks, Dolman Transportation had the choice of buying brand new, top-of-the-line, very-expensive trucks that had all sides of the hauling tank specially reinforced, or older, less-expensive trucks that only had the special reinforcement in the back.

After reviewing the relevant safety information and meeting with safety consultants, Dolman Transportation chose to buy the older, less-expensive trucks. The reason was that the brand new, top-of-the-line trucks were only safer in the very rare event that another car collided with the side of the hauling tank. In any other type of accident, the older trucks would be just as safe.

On July 16, one of Dolman Transportation's chemical-hauling trucks was hauling chemicals along a highway. Patrick Pendleton was driving in the other lane when his tire suddenly and unexpectedly blew out, causing his car to swerve and hit the side of the hauling tank on the Dolman Transportation truck.

The collision caused an explosion. The explosion left Patrick Pendleton with severe burn injuries. If the sides of the hauling tank had been specially reinforced, the explosion would not have occurred, and Patrick would not have suffered the burn injuries.

Assume the following facts are true:

Given all of the information available at the time, 90% of companies in Dolman Transportation's position would have chosen to buy the brand new, top-of-the-line chemical-hauling trucks with specially-reinforced sides.

Dolman Transportation's choice to purchase the older chemical-hauling trucks without side reinforcements led to a 1% greater risk of a side-crash-related explosion. Such an explosion would be expected to cause $5,000,000 of damage.

This means that, if it had purchased the brand new, top-of-the-line chemical-hauling trucks, Dolman Transportation would have been expected to save $50,000 of costs to other members of society. Purchasing the brand new, top-of-the-line chemical-hauling trucks would have cost Dolman Transportation $75,000 more than purchasing the older chemical-hauling trucks.

Sanders v. A & G Cosmetics

The plaintiff, Carl Sanders, is suing the defendant, A & G Cosmetics, claiming that A & G's negligence caused him injury.

Carl Sanders used Nalene, an over-the-counter baldness treatment available at drugstores. While a small amount of hair did grow back, he also had a severe adverse reaction, leaving him with permanent damage to the skin on his head and hands and a weakened immune system for life.

A & G Cosmetics is a company that sells many different cosmetic products, including wigs, "weaves," and chemical solutions designed to combat baldness. Nalene is one of the chemical solutions sold by A & G Cosmetics.

A & G Cosmetics tested Nalene extensively before putting it on the market. Based on the testing, A & G Cosmetics expected that Nalene would be effective in promoting hair growth in about 50% of customers.

In addition, based on its testing, A & G Cosmetics was aware of a very small possibility that a customer could have a severe adverse reaction—such as the one Carl had—to one of the rare chemicals in Nalene.

Before putting Nalene on the market, A & G Cosmetics could have changed its Nalene formula to an alternate formula that used different chemicals. Doing so would have reduced the risk of a severe adverse reaction. However, after reviewing the relevant manufacturing and medical information and meeting with safety and health consultants, the company decided to continue with its regular formula.

Assume the following facts are true:

A & G Cosmetics' choice to use the regular formula rather than the alternate formula led to an expected increase of three (3) cases of severe adverse reactions among Nalene users, each of which would be expected to cause $100,000 of damage.

This means that, if it had used the alternate formula for Nalene, A & G Cosmetics would have been expected to save $300,000 of costs to other members of society. Using the alternate formula for Nalene would have cost A & G Cosmetics $150,000 more than using the regular formula.

Given all of the information available at the time, 10% of companies in A & G Cosmetics' position would have switched to the alternate formula to reduce the risk of severe adverse reactions before putting their chemicals on the market.

Vignettes Used in Experiment Three

Experiment Three used the same four vignettes used in Experiments One and Two, but edited so as to provide only positive information. (Paragraphs concerning economic information were cut.)

Experiment Three also included the following fifth vignette:

Lawson v. TGI International
The plaintiff, Mary Lawson, is suing the defendant, TGI International, claiming that TGI International's negligence caused her injury.

Mary Lawson worked for years as a contractor in one of TGI International's manufacturing plants in Anytown before she developed chronic anemia. Although after a hospital stay she is now better, the condition has not fully been cured.

Mary Lawson believes that her exposure to benzene at TGI International's Anytown manufacturing plant caused her condition.

TGI International is a company that manufactures high-tech machine parts. Several years ago, the scientists at TGI International discovered that workers in the Anytown plant were often exposed to benzene, a substance that can cause anemia and leukemia.

TGI International considered buying new, state-of-the-art equipment and implementing "clean" manufacturing techniques that would have reduced workers' exposure to benzene. However, after reviewing the relevant manufacturing and medical information and meeting with safety and health consultants, the company chose to continue with its regular manufacturing technique.

Assume the following facts are true:

Given all of the information available at the time, 10% of companies in TGI International's position would have bought new, state-of-the-art equipment and implemented "clean" manufacturing techniques.

Vignettes Used in Experiment Four

Vision Case

The plaintiff, Paul Peterson, is suing the defendant, Dan Denning, claiming that Dan Denning's negligent driving caused him an injury.

On the night of March 1, Paul Peterson was walking home after going bowling at a bowling alley near his house. At a traffic light at the intersection of Mulberry Street and Sycamore Lane, Paul began crossing Mulberry Street using the pedestrian crosswalk. While crossing, Paul saw a wallet sitting in the road near the crosswalk. Paul knew the owner would want his or her wallet returned, so Paul leaned over to pick it up. But as he bent over, he lost his balance and fell forward, striking his head on the pavement, knocking him unconscious.

At the same time, Dan Denning was driving home on Mulberry Street after visiting a family member in the hospital. As Dan approached the traffic light, Dan did not see Paul lying unconscious on the road. Because the traffic light was green, Dan continued driving until he ran over the unconscious Paul's legs, causing Paul significant injury.

Assume it is a fact that, given the conditions at the time of the accident, X% of drivers in Dan's position would have seen Paul.

Hearing Case

The plaintiff, Pamela Precourt, is suing the defendant, Darla Dexter, claiming that Darla Dexter's negligence caused injuries to her young daughter, Patty Precourt.

Pamela Precourt's daughter, Patty, attended a small daycare that Darla Dexter ran from her home in Anytown. Darla Dexter was a certified childcare professional, and Pamela had always been satisfied with the care Darla provided for Patty.

Patty was in Darla's care at 2:00 p.m. on June 1, when the Anytown tornado sirens began going off. A small but rapidly-moving storm cell had just spawned a tornado on the edge of town, and it was headed for Darla's house. Darla had a windowless, interior room to which she could take the four children in her care in the event of a tornado. But inside her house with the children on June 1, she was unable to hear the sirens going off.

Because Darla did not hear the sirens, she did not move the children to the interior room. At 2:03 p.m., three minutes after the siren went off, the tornado hit Darla's home, breaking out her windows and causing other damages. Glass from a broken window cut Patty Precourt, injuring her and causing her to need an emergency surgery.

Assume it is a fact that, given Darla's location and circumstances between 2:00 and 2:03 p.m., X% of people in Darla's position would have heard the sirens.

Memory Case
The plaintiff, Priscilla Porter, is suing the defendant, Dr. Danielle Dull, claiming that Dr. Dull's negligence caused her injury.

Priscilla Porter was a patient in the Anytown Hospital, where she arrived on May 1 seeking treatment for flu-like symptoms. The evening she was admitted, Dr. Dull was trying to identify the cause of her symptoms, and after reviewing her history and doing some research, Dr Dull realized that she likely had a very dangerous and fast-acting type of infection that needed to be treated with IV antibiotics. Dr. Dull started walking down the hallway to instruct a nurse to order and administer the IV antibiotics. But before he could give the instruction, another nurse, Nurse Nancy, grabbed Dr. Dull by the arm, frantically screaming "code blue, code blue!" A patient on another ward was in deep trouble.

Dr. Dull rushed into a room with Nurse Nancy to find a patient who had seemed perfectly fine earlier in the day crashing. Over the next two hours, Dr. Dull worked frantically with the team of nurses scrambling in and out of the room and managed to stabilize the patient. When the situation was finally resolved, Dr. Dull was one hour past the scheduled end of his shift and totally frazzled. He went home and went to bed, and forgot to mention Priscilla's infection or her need for IV antibiotics to anyone.

Dr. Dull returned to the hospital the next morning and, remembering Priscilla's case, ordered the IV antibiotics. But Priscilla had gotten much worse overnight, and she ultimately had to have a leg amputated due to complications from the infection. If Priscilla had started antibiotics the previous night, she would not have lost her leg.

Assume it is a fact that, given the circumstances on the evening of May 1, X% of doctors in Dr. Dull's position would have remembered to order the IV antibiotics before the next morning.

Reaction Time Case
The plaintiff, Peter Peck, is suing the defendant, Darrell Dunn, claiming that Darrell Dunn's negligent driving caused him injury.

Peter Peck is 81 years old and no longer drives. On Sunday, June 15, Peter attended services at the church across the street from his house, as was his custom. After church, Darrell Dunn, a friend of Peter's son, picked Peter up to drive him to a physical therapy appointment.

The physical therapists' office was located 15 miles up the highway from Peter's house. Peter rested in the passenger seat as Darrell drove along the highway, obeying the 70 mile-per-hour speed limit.

10 miles into the drive, a deer darted out onto the highway in front of Darrell's vehicle. Darrell was momentarily startled, but as soon as he processed what was happening, he slammed the brakes on his car.

Unfortunately, the car did not stop before it hit the deer. After the collision, the injured deer jumped onto the windshield, shattering it and causing significant injury to Peter.

Assume it is a fact that, given the conditions at the time of the accident, X% of drivers in Darrell's position would have been able to react (i.e. brake) in time to stop the car before it hit the deer.

Decision Making Case

The plaintiff, Patrick Pendleton, is suing the defendant, Donald Dolman, claiming that Donald Dolman's negligence caused him injury.

After years working as an accountant, Donald Dolman decided he wanted to change careers and travel more. Donald started his own trucking company, which focused on hauling chemicals over long distances.

Donald needed to buy a special truck for hauling chemicals. He could either buy a brand new, top-of-the-line, very-expensive truck that had all sides of the hauling tank specially reinforced, or an older, less expensive truck that only had the special reinforcement in the back.

After reviewing all of the information he could find and speaking with some safety consultants, Donald Dolman chose the older, less expensive truck. Donald reasoned that the brand new, top-of-the-line truck would only be safer in the very rare event that another car collided with the side of the hauling tank. In any other type of accident, the older truck would be just as safe.

On July 16, Donald Dolman was hauling chemicals along a highway in his truck. Patrick Pendleton was driving in the other lane when his tire suddenly and unexpectedly blew out, causing his car to swerve and hit the side of Donald's hauling tank.

The collision caused an explosion. The explosion left Patrick Pendleton with severe burn injuries. If the sides of the hauling tank had been specially reinforced, the explosion would not have occurred, and Patrick would not have suffered the burn injuries.

Assume it is a fact that, given all of the information available at the time, X% of people in Donald Dolman's position would have chosen to buy the brand new, top-of-the-line truck with specially-reinforced sides.

Jury Instructions Used in Experiments

The judge asks you to decide whether the defendant, [Defendant's Name], was negligent.

In connection with this question, the judge provides the following instructions:

This case involves claims of negligence. Negligence is the lack of ordinary care; that is, the absence of the kind of care a reasonably prudent and careful person would exercise in similar circumstances. That standard is your guide. If a person's conduct in a given circumstance doesn't measure up to the conduct of an ordinarily prudent and careful person, then that person was negligent. On the other hand, if the person's conduct does measure up to the conduct of a reasonably prudent and careful person, the person wasn't negligent.

The mere fact that an accident occurred isn't enough to establish negligence.

Contingency Tables Summarizing Negligence Verdicts by Condition

Experiment One

	Positive Information – 10% Take Precaution	Positive Information – 90% Take Precaution	**Row Total**
Economic Information – Cost-Justified (B<PL)	55 of 99	76 of 99	131 of 198
Economic Information – Non-Cost-Justified (B>PL)	42 of 99	76 of 99	118 of 198
Column Total	97 of 198	152 of 198	**249 of 396 (Grand Tot)**

Table 1.7. Number of participants finding defendant negligent in each condition in Experiment One. Note that each participant responded to four cases, one in each cell.

Experiment Two

	Positive Information – 10% Take Precaution	Positive Information – 90% Take Precaution	**Row Total**
Economic Information – Cost-Justified (B<PL)	57 of 98	74 of 98	131 of 196
Economic Information – Non-Cost-Justified (B>PL)	53 of 98	78 of 98	131 of 196
Column Total	110 of 196	152 of 196	**262 of 392 (Grand Tot)**

Table 1.8. Number of participants finding defendant negligent in each condition in Experiment Two. Note that each participant responded to four cases, one in each cell.

Experiment Three

PPP Condition	# Finding Defendant Negligent
0	29 of 60
10	34 of 60
25	35 of 60
50	44 of 60
90	49 of 60
Grand Total	**191 of 300**

Table 1.9. Number of participants finding defendant negligent in each condition in Experiment Three. Note that each participant responded to five cases, one in each cell.

Experiment Four

PPP Level	# Finding Defendant Negligent
0	15 of 53
10	15 of 53
25	20 of 53
50	24 of 53
90	39 of 53
Grand Total	**113 of 265**

Table 1.10. Number of participants finding defendant negligent in each condition in Experiment Four. Note that each participant responded to five cases, one in each cell.

Appendix B. Appendix to Chapter 2.

Stimuli for Experiment 1 (fonts as presented to participants)

Story:

Uber's self-driving car killed someone. What happened?

by Nathaniel Meyersohn and Matt McFarland @CNNMoney
March 20, 2018: 9:42 AM ET

A self-driving Uber SUV struck and killed a pedestrian on a street Sunday in Tempe, Arizona. It's believed to be the first fatality involving a fully autonomous car. Here's what you need to know about the crash, Uber's autonomous vehicle testing, and what's next.

What happened?

A self-driving Uber Volvo XC90 SUV killed 49-year-old Elaine Herzberg as she walked her bicycle across a street in Tempe, Arizona, Sunday night, according to the Tempe Police Department.

Based on preliminary information, the car was going approximately 40 mph in a 35 mph zone, according to Tempe Police Detective Lily Duran.

Rafael Vasquez, a 44-year-old test driver from Uber, was behind the wheel of the car at the time.

The department is investigating the crash.

Police say the investigation so far does not indicate the SUV slowed down before hitting Herzberg.

Why was there a driver behind the wheel?

The car was in autonomous mode at the time of the crash, police said.

Autonomous mode means the car is driving on its own. During tests, a person sits behind the wheel as a safeguard.

There were no signs Vasquez was impaired after the collision, Sgt. Ronald Elcock, a Tempe

police spokesman, said in a press conference.

Has Uber responded?

Uber said it has stopped testing self-driving vehicles throughout the United States and Canada. It's currently conducting autonomous vehicle tests in Arizona, Pittsburgh, Toronto and other areas.

In a statement, Uber said it is "fully cooperating" with local officials. "Our hearts go out to the victim's family," the company said in a statement. CEO Dara Khosrowshahi reacted on Twitter.

Uber has previously grounded its vehicles while investigating a crash. In 2017, Uber briefly pulled its vehicles from roads after an Uber self-driving vehicle in Tempe landed on its side.

Which authorities are investigating the crash?

In addition to the Tempe Police Department, the National Transportation Safety Board said it is launching an investigation.

The Maricopa County Attorney's Office will ultimately determine whether any charges will be filed in the crash.

Why did this happen in Arizona?

Arizona is an incubator of self-driving car tests.

Earlier this month, Arizona Governor Doug Ducey updated an executive order to allow self-driving cars to drive on state roads without a test driver behind the wheel.

Arizona has little inclement weather. This makes it more appealing for self-driving cars, which have less experience in rain or during snowfall.

Who else is testing self-driving cars?

Waymo, the self-driving arm of Google's (GOOG) parent company, is launching a public self-driving car service this year in the Phoenix, Arizona, area. Companies such as GM's (GM) Cruise and Intel (INTC) are also testing in the state.

Many self-driving companies have circled 2020 as the date when self-driving vehicle technology would be deployed on American roads.

Additional Information (viewed only in Additional Information condition):

Additional Information

Before answering additional questions, please read the following additional information and assume, for the purposes of this study, that it is completely accurate:

Investigation of the accident (including a review of video of the accident) reveals that the pedestrian who was killed stepped out in front of the oncoming vehicle, even though the vehicle was plainly visible from the pedestrian's position.

There was no time for the vehicle to stop in response to the pedestrian entering the roadway. In fact, **no human driver** could have possibly responded fast enough to brake before colliding with the pedestrian.

Even if the technology had somehow applied the breaks instantaneously, the car would not have slowed down significantly before striking the pedestrian. For these reasons, investigators conclude that the accident was "unavoidable" once the pedestrian entered the roadway.

Stimuli for Experiments 2 and 3 (fonts as presented to participants)

Uber Story:

Anthropomorphic Framing Version

Uber's self-driving car killed someone. What happened?

A self-driving Uber SUV struck and killed a pedestrian on a street Sunday in Tempe, Arizona. The self-driving SUV, named Sporty, is believed to be the first fully autonomous vehicle to kill a human. Here's what you need to know about the crash, about Sporty, and about what's next.

What happened?

Sporty was driving down a thoroughfare in Tempe, Arizona, on Sunday night, when Sporty struck and killed a 49-year-old woman as she walked her bicycle across the street, according to the Tempe Police Department.

Based on preliminary information, Sporty was travelling approximately 40 mph in a 35 mph zone, according to police.

Carl Jackson, a 44-year-old test driver from Uber, was behind the wheel in Sporty, but Sporty was driving autonomously.

It does not appear that Sporty slowed down before hitting the pedestrian.

The investigation suggests that, although the pedestrian was in Sporty's line of sight, Sporty's attention was focused elsewhere. Sporty did not pay attention to the pedestrian.

Why was there a driver behind the wheel?

Sporty was operating autonomously at the time of the crash, police said.

This means that Sporty was making all of its own decisions, driving entirely on its own. During autonomous tests, a person sits behind the wheel as a safeguard.

There were no signs that the test driver, Carl Jackson, was impaired at the time of the collision.

Why did this happen in Arizona?

Arizona is an incubator of self-driving car tests.

Recently, Arizona's Governor updated an executive order to allow self-driving cars to drive on state roads without a test driver behind the wheel.

Arizona has little inclement weather. This makes it more appealing for self-driving cars, which have less experience driving in rain or during snowfall.

Who else is testing self-driving cars?

Waymo, the self-driving arm of Google's (GOOG) parent company, is launching a public self-driving car service this year in the Phoenix, Arizona, area. Companies such as GM's (GM) Cruise and Intel (INTC) are also testing in the state.

Many self-driving companies have circled 2020 as the date when self-driving vehicle technology would be deployed on American roads.

Mechanical Framing Version

Uber self-driving car malfunction caused a fatality. What happened?

A pedestrian was struck and killed by a self-driving Uber Volvo XC90 SUV on a street Sunday in Tempe, Arizona. It's believed to be the first fatality involving a car functioning fully autonomously. Here's what you need to know about the crash, about the autonomous vehicle, and what's next.

What happened?

A self-driving Uber Volvo XC90 SUV malfunctioned on a thoroughfare in Tempe, Arizona, on Sunday night, colliding with a 49-year-old woman as she walked her bicycle across the street, according to the Tempe Police Department. The pedestrian died in the collision.

Based on preliminary information, the car was traveling at approximately 40 mph in a 35 mph zone, according to police.

Carl Jackson, a 44-year-old test driver from Uber, was behind the wheel of the car at the time.

It does not appear that the car's central processor initiated the braking protocol before hitting the pedestrian.

The investigation suggests that, although the car's photosensors picked up the signal of the woman crossing the street in front of the car, the data from the photosensors was not processed. The car's processor was processing input from other sensors at the time.

Why was there a driver behind the wheel?

The car was in autonomous mode at the time of the crash, police said.

Autonomous mode means the car is being navigated entirely by its onboard computers. During tests of autonomous mode, a person sits behind the wheel as a safeguard.

There were no signs that the test driver, Carl Jackson, was impaired at the time of the collision.

Why did this happen in Arizona?

Arizona is an incubator of self-driving car tests.

Recently, Arizona's Governor updated an executive order to allow manufacturers to test self-driving cars on state roads without a test driver behind the wheel.

Arizona has little inclement weather. This makes it an appealing place to test self-driving cars, as testers can avoid complications related to rain and snowfall.

Who else is testing self-driving cars?

Waymo, the self-driving arm of Google's (GOOG) parent company, is launching a public self-driving car service this year in the Phoenix, Arizona, area. Companies such as GM's (GM) Cruise and Intel (INTC) are also testing in the state.

Many self-driving companies have circled 2020 as the date when self-driving vehicle technology would be deployed on American roads.

Tesla Story:

Anthropomorphic Framing Version

Who's to blame when self-driving cars crash?

On May 7, 2016, Tess made history. Tess is the name given to Tesla's pilot self-driving car. On May 7, Tess became the first self-driving car to kill its driver, Carl Jackson.

Jackson had turned on Tess's autonomous driving system, meaning that Tess was making all of the decisions at the time of the accident, driving itself. Tess set the cruise control at 74 miles per hour. As Tess sped down a Florida highway, a tractor-trailer came out of an intersecting road.

While driving, Tess keeps its eyes on the road in front of it, scanning for potential hazards. But on this day, Tess did not recognize the tractor-trailer crossing the highway, because the color and reflection from the tractor-trailer blended in perfectly with the bright sky behind it. Because Tess did not see the tractor-trailer, Tress did not brake, nor did Tess issue any warning to the driver. Tess crashed into the trailer, killing Jackson.

No one knows for sure what Jackson was doing in the last seconds of his life. But the driver of the tractor-trailer told police he heard the sounds from a Harry Potter movie playing in Tess immediately after the crash.

In investigating the crash, the National Highway Traffic Safety Administration stressed that Tesla's autonomous driving system was intended to aid, not replace, human drivers. The technology, however, is changing. Google, Mercedes-Benz, Tesla, Uber and Volvo are some of the companies working to develop fully autonomous cars, intended to drive themselves without allowing for human intervention. Some of Google's prototypes don't have steering wheels or brake pedals.

Mechanical Framing Version

Who's to blame when self-driving cars malfunction?

On May 7, 2016, Carl Jackson made history. The Ohio resident became the first person to die as the result of a self-driving car malfunction.

Jackson had turned on Tesla's autonomous driving system in his Tesla Model S, meaning that the car was being navigated entirely by its onboard computer. The computer set the cruise control to 74 miles per hour, and as the computer steered the car swiftly down a Florida highway, a tractor-trailer came out of an intersecting road.

While the Tesla is operating in autonomous mode, it uses onboard radar and cameras to scan for potential hazards, sending data to its central computer. But on this day, the central computer did not detect the tractor-trailer crossing the highway, because the input it received from the white tractor-trailer was so similar to the input it received from the bright sky behind it. Because the computer did not detect the tractor-trailer, it did not initiate the braking protocol, nor did it issue any warning to the driver. The car collided with the trailer, killing Jackson.

No one knows for sure what Jackson was doing in the last seconds of his life. But the driver of the tractor-trailer told police he heard the sounds from a Harry Potter movie playing in the crushed Tesla Model S immediately after the crash.

In investigating the crash, the National Highway Traffic Safety Administration stressed that Tesla's autonomous driving system was intended to aid, not replace, human drivers. The technology, however, is changing. Google, Mercedes-Benz, Tesla, Uber and Volvo are some of the companies working to develop fully autonomous cars, intended to be driven entirely by onboard computers without human intervention. Some of Google's prototypes don't have steering wheels or brake pedals.

REFERENCES

Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of Consumer Research*, *34*, 468–479.

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind?. *Neuroimage*, *30*(3), 1059-1068.

American Law Institute. (1965). *Restatement of the Law (Second), Torts*. St. Paul, MN: America Law Institute Publishers.

American Law Institute. (2010). *Restatement of the Law (Third), Torts, Liability for Physical and Emotional Harm*. St. Paul, MN: America Law Institute Publishers.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?.*Psychological review*, *116*(4), 953-970.

Austin, R. T. (1992). Better Off with the Reasonable Man Dead or the Reasonable Man Did the Darndest Things. *BYU L. Rev.*, 479-492.

Baddeley, A. D., Grant, W., Wight, E., & Thomson, N. (1975). Imagery and visual working memory. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 205–217). London, UK: Academic Press.

Bago, B., & De Neys, W. (2019). The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257-299.

Baker, L. J., Hymel, A. M., & Levin, D. T. (2018). Anthropomorphism and intentionality improve memory for events. *Discourse Processes*, *55*(3), 241-255.

Baker, L. J., Jaeger, C. B., Havard, A., Lane, J. D., Harriott, C., Adams, J., & Levin, D. T. Cognitive dissonance increases attributions of agency to intelligent agents (in prep).

Baker, L. J., Levin, D. T., & Saylor, M. M. (2016). The extent of default visual perspective taking in complex layouts. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(4), 508.

Baltimore & Ohio R. Co. v. Goodman, 275 U.S. 66 (1927).

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. *Cognitive psychology*, *31*(3), 219-247.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241-251.

Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, *9*(3), 415-433.

Bartz, J. A., Tchalova, K., & Fenerci, C. (2016). Reminders of social connection can attenuate anthropomorphism: A replication and extension of Epley, Akalis, Waytz, and Cacioppo (2008). *Psychological Science*, *27*(12), 1644-1650.

Bigelow, A. E., & Dugas, K. (2009). Relations among preschool children's understanding of visual perspective taking, false belief, and lying. *Journal of Cognition and Development*, *9*(4), 411-433.

Bragan ex. Rel. Bragan v. Symanzik, 687 N.W.2d 881 (Mich. Ct. App. 2004).

Brown-Schmidt, S., & Heller, D. (2018). Perspective-taking during conversation. In S. Rueschemeyer & M. G. Gaskell (Eds.), *Oxford handbook of psycholinguistics* (pp. 551–574). New York: Oxford University Press.

Burch v. American Family Mutual Insurance Co., 543 N.W.2d 282 (Wis. 1996).

Burckitt, B. (2018, June 22). Fatal Uber crash: A timeline of the crash and investigation. *Arizona Republic*. Retrieved from https://www.azcentral.com/story/news/local/tempe/2018/06/22/fatal-uber-crash-timeline-crash-and-investigation/725921002/.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606-637.

Calo, R. (2014, Sept. 15). The case for a federal robotics commission. *Brookings Institute*. Retrieved from https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission/.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*, 513-563.

Cane, J. E., Ferguson, H. J., & Apperly, I. A. (2017). Using perspective to resolve reference: The impact of cognitive load and motivation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 591.

Carruthers, P. (2016). Two systems for mindreading?. *Review of Philosophy and Psychology*, *7*(1), 141-162.

Carruthers, P. (2017). Mindreading in adults: evaluating two-systems views. *Synthese*, *194*(3), 673-688.

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*, 48-64.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology* (Vol. 9, pp. 287-299). Amsterdam: North-Holland.

Collins, R. K. (1976). Language, history and the legal process: a profile of the reasonable man. *Rutgers-Camden Law Journal*, *8*, 311-324.

Cosgrove, I. J. (2015). The Illusive Reasonable Person: Can Neuroscience Help the Mentally Disabled. *Notre Dame Law Review*, *91*, 421-446.

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, *1*(2), 255-259.

de Melo, C. M., Gratch, J., & Carnevale, P. J. (2014). The Importance of Cognition and Affect for Artificially Intelligent Decision Makers. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 336-342). AAAI.

de Melo, C. M., Gratch, J., & Carnevale, P. J. (2015). Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing*, *6*(2), 127-136.

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental psychology*, *52*(5), 758-771.

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002, June). All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 321-326). ACM.

Eggen, J. M., & Laury, E. J. (2011). Toward a neuroscience model of tort law: How functional neuroimaging will transform tort doctrine. *Columbia Science & Technology Law Review*, *13*, 235-306.

Elekes, F., Varga, M., & Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, *41*, 93-103.

Elekes, F., Varga, M., & Király, I. (2017). Level-2 perspectives computed quickly and spontaneously: Evidence from eight-to 9.5-year-old children. *British Journal of Developmental Psychology*, *35*(4), 609-622.

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological science*, *19*(2), 114-120.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, *87*(3), 327-339.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social cognition*, *26*(2), 143-155.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864-886.

Feigenson, N. R. (1995). The rhetoric of torts: How advocates help jurors think about causation, reasonableness, and responsibility. *Hastings Law Journal*, 47, 61-165.

Feldman, H. L. (1998) Prudence, Benevolence, and Negligence: Virtue Ethics and Tort Law, *Chicago-Kent Law Review*, *74*, 1431-1466.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology*, *17*(1), 99-103.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition*, *44*(3), 283-296.

Fogg, B. J., & Nass, C. (1997). Silicon sycophants: the effects of computers that flatter. *International journal of human-computer studies*, *46*(5), 551-561.

Fleming, J. G. (1971). *The law of torts (4th ed.)* Sydney: The Law Book Co.

Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008, March). How people anthropomorphize robots. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 145-152). IEEE.

Goldberg, J. C. P., Sebok, A. J., & Zipursky, B. C. (2004). *Tort law: Responsibilities and redress*. New York: Aspen.

Goldberg, J. C., & Zipursky, B. C. (2009). Torts as wrongs. *Tex. L. Rev.*, *88*, 917.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational psychologist*, *40*(4), 225-234.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *science*, *315*(5812), 619-619.

Hans, V. P. (1995). The contested role of the civil jury in business litigation. *Judicature*, *79*, 242-248.

Hans, V. P., & Ermann, M. D. (1989). Responses to corporate versus individual wrongdoing. *Law and Human Behavior*, *13*(2), 151-166.

Harley, E. M. (2007). Hindsight bias in legal decision making. *Social Cognition*, *25*(1), 48-63.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243-259.

Henderson Jr, J. A. (1975). Expanding the negligence concept: Retreat from the rule of law. *Indiana Law Journal*, *51*, 467-527.

Herbert, A. P. (1930). *Misleading cases in the common law.* London: Methuen.

Hetcher, S. (2001). Non-utilitarian negligence norms and the reasonable person standard. *Vand. L. Rev.*, *54*, 863-892.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground?. *Cognition*, *59*(1), 91-117.

Hurd, H. M., & Moore, M. S. (2002). Negligence in the Air. *Theoretical Inquiries in Law*, *3*(2), 333-412.

Hymel, A. M., Levin, D. T., Barrett, J., Saylor, M., & Biswas, G. (2011, July). The Interaction of Children's Concepts about Agents and Their Ability to Use an Agent-Based Tutoring System. In *International Conference on Human-Computer Interaction* (pp. 580-589). Springer, Berlin, Heidelberg.

Jaeger, C. B., Hymel, A. M., Levin, D. T., Biswas, G., Paul, N., & Kinnebrew, J. (2019). The interrelationship between concepts about agency and students' use of teachable-agent learning technology. *Cognitive research: principles and implications*, *4*(1), 14.

Jaeger, C. B., & Levin, D. T. (2016). If asimo thinks, does roomba feel?: The legal implications of attributing agency to technology. *Journal of Human-Robot Interaction*, *5*(3), 3-25.

Jaeger, C. B., & Levin, D. T. (2017). "The car wrecked me": How concepts of agency affect negligence decisions. Presented at Conference on Empirical Legal Studies, Ithaca, NY.

Jaeger, C. B., Levin, D. T., & Porter, E. (2017). Justice is (change) blind: Applying research on visual metacognition in legal settings. *Psychology, Public Policy, and Law*, *23*(2), 259.

James Jr, F. (1951). The Qualities of the Reasonable Man in Negligence Cases. *Missouri Law Review*, *16*, 1-26.

Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 549-559.

Jones, C. R., Simonoff, E., Baird, G., Pickles, A., Marsden, A. J., Tregay, J., ... & Charman, T. (2018). The association between theory of mind, executive function, and the symptoms of autism spectrum disorder. *Autism Research*, *11*(1), 95-109.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... & Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology*, *48*(2), 303-314.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin,& D. Kahneman (Eds.),*Heuristics and biases* (pp. 49 –81). New York: Cambridge University Press.

Kahneman, D., Schkade, D., & Sunstein, C. (1998). Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty*, *16*(1), 49-86.

Kelley, P. J., & Wendt, L. A. (2001). What judges tell juries about negligence: A review of pattern jury instructions. *Chi.-Kent L. Rev.*, *77*, 587-682.

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition*, *70*(3), 241-272.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32-38.

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, *39*(1), 1-20.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25-41.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: an ineradicable and egocentric bias in social perception. *Journal of personality and social psychology*, *67*(4), 596-610.

Lane, J. D., Wellman, H. M., & Evans, E. M. (2010). Children's understanding of ordinary and extraordinary minds. *Child development*, *81*(5), 1475-1489.

Lane, J. D., Evans, E. M., Brink, K. A., & Wellman, H. M. (2016). Developing concepts of ordinary and extraordinary communication. *Developmental psychology*, *52*(1), 19-30.

Lecce, S., Bianco, F., Devine, R. T., & Hughes, C. (2017). Relations between theory of mind and executive function in middle childhood: A short-term longitudinal study. *Journal of experimental child psychology*, *163*, 69-86.

Lehman, B., D'Mello, S., & Graesser, A. (2012). Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, *15*(3), 184-194.

Leslie, A. (1994a). ToMM, ToBY, and Agency: Core Architecture and Domain Specificity. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the Mind: domain specificity in cognition and culture*. Cambridge, UK: Cambridge University Press.

Levin, D. T., Adams, J. A., Saylor, M. M., & Biswas, G. (2013). A transition model for cognitions about agency. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction* (pp. 373-380). IEEE Press.

Levin, D. T., & Angelone, B. L. (2008). The visual metacognition questionnaire: A measure of intuitions about vision. *The American Journal of Psychology*, *121*(3), 451-472.

Levin, D. T., Harriott, C., Paul, N. A., Zhang, T., & Adams, J. A. (2013). Cognitive dissonance as a measure of reactions to human-robot interaction. *Journal of Human-Robot Interaction*, *2*(3), 3-17.

Levin, D. T., Killingsworth, S. S., & Saylor, M. M. (2008, March). Concepts about the capabilities of computers and robots: A test of the scope of adults' theory of mind. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 57-63). IEEE.

Levin, D. T., Killingsworth, S. S., Saylor, M. M., Gordon, S. M., & Kawamura, K. (2013). Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human–Computer Interaction*, *28*(2), 161-191.

Levin, D. T., Saylor, M. M., & Lynn, S. D. (2012). Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers. *International Journal of Human-Computer Studies*, *70*(8), 527-534.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence*, *175*(5-6), 942.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551-556.

Loewenstein, G., O'Donoghue, T., & Bhatia, S. (2015). Modeling the interplay between affect and deliberation. *Decision*, *2*(2), 55.

MacCoun, R. J. (1996). Differential Treatment of Corporate Defendants by Juries: An Examination of the Deep-Pocket Hypothesis. *Law & Society Review*, *30*, 121-162.

Mainwaring, S. D., Tversky, B., Ohgishi, M., & Schiano, D. J. (2003). Descriptions of simple spatial scenes in English and Japanese. *Spatial cognition and computation*, *3*(1), 3-42.

Majdandžić, J., Bauer, H., Windischberger, C., Moser, E., Engl, E., & Lamm, C. (2012). The human factor: behavioral and neural correlates of humanized perception in moral decision making. *PloS one*, *7*(10), e47698.

Maldonado, T. (2016). *The role of working memory capacity and cognitive load in producing and detecting deception* (Doctoral dissertation, Montana State University-Bozeman, College of Letters & Science).

Martin, J., & Sokol, B. (2011). Generalized others and imaginary audiences: A neo-Meadian approach to adolescent egocentrism. *New Ideas in Psychology*, *29*(3), 364-375.

May, F., & Monga, A. (2013). When time has a will of its own, the powerless don't have the will to wait: Anthropomorphism of time can decrease patience. *Journal of Consumer Research*, *40*(5), 924-942.

Mead, G. H. (1934). *Mind, Self, and Society from the Standpoint of a Social Behaviorist*. Chicago: University of Chicago Press.

Meltzer BN. (1972). Mead's social psychology. In: JG Manis, BN Meltzer (eds) *Symbolic Interaction: A Reader in Social Psychology (2nd ed.)* (pp. 4-22). Boston: Allyn and Bacon, 1972.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & psychophysics*, *68*(2), 327-337.

Miller, A. D., & Perry, R. (2012). The reasonable person. *New York University Law Review*, *87*, 323-392.

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication research*, *23*(6), 651-674.

Murdock, C. W., & Sullivan, B. (2012). What Kahneman Means for Lawyers: Some Reflections on Thinking, Fast and Slow. *Loy. U. Chi. LJ*, *44*, 1377.

Meyersohn, N. & McFarland, M. (2018, March 20). Uber's self-driving car killed someone. What happened? *CNN Money*. Retrieved from https://money.cnn.com/2018/03/20/news/companies/self-driving-uber-death/index.html.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates?, *International Journal of Human-Computer Studies*, *45*(6), 669-678.

Nass, C., Isbister, K., & Lee, E. (2000). Truth is beauty: Researchingembodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 374 – 402).Cambridge, MA: MIT Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81-103.

Nass, C., Moon, Y., & Carney, P. (1999). Are People Polite to Computers? Responses to Computer-Based Interviewing Systems 1. *Journal of Applied Social Psychology*, *29*(5), 1093-1109.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995). Can computer personalities be human personalities?. In *Conference companion on Human factors in computing systems* (pp. 228-229). ACM.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, *27*(10), 864-876.

Nickerson, R. S. (2001). The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Science*, *10*(5), 168-172.

Noll v. Marian, 32 A.2d 18 (Pa. 1943).

Osborne v. Montgomery, 234 N.W. 372 (Wis. 1931).

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and brain sciences*, *1*(4), 515-526.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, *117*(2), 230-236.

Rachlinski, J. J. (1998). A positive psychological theory of judging in hindsight. *The University of Chicago Law Review*, *65*(2), 571-625.

Rachlinski, J. J. (2002). Misunderstanding ability, misallocating responsibility. *Brook. L. Rev.*, *68*, 1055-1092.

Raver, C., & Leadbeater, B. J. (1993). The problem of the other in research on theory of mind and social development. *Human Development*, *36*(6), 350-362.

Richards, N. M., & Smart, W. D. (2016). How should the law think about robots?. In *Robot law*. Northampton, MA: Edward Elgar Publishing.

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical perspectives*, *18*, 295-313.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*(5), 898-915.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255-1266.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755-1758.

Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*(10), 1360-1370.

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological science*, *23*(8), 842-847.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*(1), 1-24.

Seidenberg, S. (2017, July 1). Who's to blame when self-driving cars crash? *ABA Journal*. Retrieved from http://www.abajournal.com/magazine/article/selfdriving_liability_highly_automated_vehicle.

Sevier, J. (2014). Testing Tribe's Triangle: Juries, Hearsay, and Psychological Distance. *Georgetown Law Journal*, *103*, 879-932.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401-411.

Siino, R. M., Chung, J., & Hinds, P. J. (2008, August). Colleague vs. tool: Effects of disclosure in human-robot collaboration. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 558-562). IEEE.

Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *University of Chicago Law Review*, *71*, 511-586.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3-22.

Sodian, B. (1991). The development of deception in young children. *British journal of developmental psychology*, *9*(1), 173-188.

Sood, A. M. (2014). Cognitive Cleansing: Experimental Psychology and the Exclusionary Rule. *Georgetown Law Journal*, *103*, 1543-1608.

Sood, A. M. (2019). Attempted Justice: Misunderstanding and Bias in Psychological Constructions of Critical Attempt. *Stanford Law Review*, *71*, 593-686.

Strait, M., Briggs, G., & Scheutz, M. (2013). Some correlates of agency ascription and emotional value and their effects on decision-making. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference* (pp. 505-510). IEEE.

Surtees, A. D., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child development*, *83*(2), 452-460.

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*(2), 426-438.

Surtees, A. D., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology*, *30*(1), 75-86.

Surtees, A., Samson, D., & Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97-105.

Todd, A. R., Forstmann, M., Burgmer, P., Brooks, A. W., & Galinsky, A. D. (2015). Anxious and egocentric: How specific emotions influence perspective taking. *Journal of Experimental Psychology: General*, *144*(2), 374-391.

Todd, A. R., & Simpson, A. J. (2016). Anxiety impairs spontaneous perspective calculation: Evidence from a level-1 visual perspective-taking task. *Cognition, 156,* 88-94.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675-691.

Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, *110*(1), 124-129.

United States v. Carroll Towing Co., 159 F.2d 169 (2d Cir. 1947).

Vaughan v. Menlove, 132 Eng. Rep. 490 (C. P. 1837).

Vladeck, D. C. (2014). Machines without principles: Liability rules and artificial intelligence. *Washington Law Review*, *89*, 117-150.

Votruba, A. M. (2013). Will the Real Reasonable Person Please Stand Up: Using Psychology to Better Understand How Juries Interpret and Apply the Reasonable Person Standard. *Arizona State Law Journal*, *45*, 703-732.

Vytal, K., Cornwell, B., Arkin, N., & Grillon, C. (2012). Describing the interplay between anxiety and cognition: from impaired performance under low cognitive load to reduced anxiety under high load. *Psychophysiology*, *49*(6), 842-852.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219-232.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of personality and social psychology*, *99*(3), 410-435.

Westra, E. (2017). Spontaneous mindreading: A problem for the two-systems account. *Synthese*, *194*(11), 4559-4581.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Woodman, G. F., Vogel, E. K., & Luck, S. J. (2001). Visual search remains efficient when visual working memory is full. *Psychological Science*, *12*(3), 219-224.

Yuan, L., & Dennis, A. (2017). Interacting like humans? Understanding the effect of anthropomorphism on consumer's willingness to pay in online auctions. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.