2018

Trust in Humans and Robots: Economically Similar but Emotionally Different

Eric Schniter
Chapman University, schniter@chapman.edu

Timothy W. Shields
Chapman University, shields@chapman.edu

Daniel Sznycer
University of Montreal

Recommended Citation

Schniter, E., Shields, T. W., & Sznycer, D. (2019). Trust in humans and robots: Economically similar but emotionally different. ESI Working Paper 18-22. Retrieved from https://digitalcommons.chapman.edu/esi_working_papers/282/

Title

Trust in Humans and Robots: Economically Similar but Emotionally Different.

Authors

E. Schniter, [1,2]* T. W. Shields,[1,2] D. Sznycer[3]

Affiliations

[1] Economic Science Institute, Chapman University, One University Drive, Orange, CA 92866, USA.

[2] Argyros School of Business and Economics, Chapman University, One University Drive, Orange, CA 92866, USA.

[3] Department of Psychology, University of Montreal, Pavillon Marie-Victorin, 90 Vincent d'Indy Ave., Montreal, QC, H3C 3J7, Canada.

*Corresponding author email: eschniter@gmail.com

WORKING PAPER UNDER REVIEW– PLEASE DO NO CIRCULATE WITHOUT AUTHORS PERMISSION

Abstract Trust-based interactions with robots are increasingly common in the marketplace, workplace, on the road, and in the home. However, a looming concern is that people may not trust robots as they do humans. While trust in fellow humans has been studied extensively, little is known about how people extend trust to robots. Here we compare trust-based investments and emotions from across three nearly identical economic games: human-human trust games, human-robot trust games, and human-robot trust games where the robot decision impacts another human. Robots in our experiment mimic humans: they are programmed to make reciprocity decisions based on previously observed behaviors by humans in analogous situations. We find that people invest similarly in humans and robots. By contrast, the social emotions elicited by the interactions (but not non-social emotions) differed across human and robot trust games, and did so lawfully. Emotional reactions depended on how one's trust game decision interacted with the partnered agent's decision, and whether another person was affected economically and emotionally.

Keywords: Trust, Robots, Artificial Intelligence, Emotion, Experiment

Introduction

Trust-based interactions between humans and robots are increasingly common in healthcare, the marketplace, at home, and on the road (Lee, Knox, Baumann, Breazeal, & DeSteno, 2013; Schaefer, Chen, Szalma, & Hancock, 2016). People are more likely to interact with automation, artificial intelligence (AI), and robots that they can trust (Lee & See, 2004). However, a psychology that evolved to navigate interactions with *fellow humans* regulates this trust, and people worry that we may overly-trust robots (Robinette, Li, Allen, Howard, & Wagner, 2016; Salem, Lakatos, Amirabdollahian, & Dautenhahn, 2015). But if over-trusting robots is a problem, so is under-trusting benign robots and thus foregoing various benefits. Previous work shows that emotions calibrate trust when people interact with fellow humans (Schniter & Sheremeta, 2014). Whether emotions affect trust-based interactions with robots remains an open question. Some view robots as not capable of being "socially and psychologically present" in terms of benevolence, other-regard, and integrity (Gefen & Straub, 2004). For example, in their interactions with automated online vendors, people perceive them as "agentic" (e.g. capable of reliable response) but lacking in "experiential" mental states such as having desires, feelings, and emotions (Rai & Diermeier, 2015). Further, trust could also be affected by whether robots act alone or on behalf of other humans whom they affect. As robots and AI increasingly replace human agency, we find ourselves in interactions with non-human robotic agents and automated systems that can affect not only us, but also the welfare of other humans they serve (Clark, 2003; Groom & Nass, 2007).

Despite the problems identified with trust in autonomous systems, there is little agreement over *how much trust* is appropriate or otherwise expected (e.g., how much do people normally trust one another and robots, and how much does this vary?). Across a number of studies, researchers have shown that humans initially act towards computers and robots as they do towards humans, applying social norms (Rai & Diermeier, 2015). Meanwhile, "uncanny valley" phenomena challenges this equivalence by highlighting people's distrust and revulsion when interacting with virtual agents that closely resemble humans, rather than the trust and empathy they feel with fellow humans (Mathur & Reichling, 2016). Conflicting results may stem from inconsistent and inappropriate research methods. Hancock et al.'s (2011) meta-analysis of trust in human-robot interactions noted with concern that current knowledge of trust in robots is derived almost exclusively from subjective responses, rather than more objective methods such as incentive-compatible economic trust games. Deception of participants by experimenters is also common across human-robot interaction studies and may contribute to unreliable responses. The concern is over discrepancies between subjective response and behavior: for example, while people report they trust a robot, their behavior shows mistrust.

Here we report the results of an anonymous trust-game experiment. In it, a human investor first decides how much of a $10 endowment to entrust to a trustee. The experimenter multiplies the entrusted amount by three - creating potential gains from trust. Then the trustee receives this and decides how much to reciprocate to the investor. The standard game theoretic (Nash equilibrium) prediction is that trustees will keep everything that is entrusted to them. Further, investors are predicted to anticipate this and thus entrust nothing. In trust games, however, most people demonstrate trust and reciprocation such that, often, both investors and trustees benefit (Johnson & Mislin, 2011).

In our experiment, human investors interact with one of three types of trustees: a fellow human, a robot, or a robot whose payoffs go to another human. We programmed robot trustees to mimic previously observed reciprocation by human trustees. After the trust-game interaction, participants rate how much they feel various positive and negative emotions. Our experimental design allows us to illuminate three important aspects of trust in robots. First, we can determine how much humans trust robots compared to fellow humans. To do this we rely on versions of the "trust game" (Berg, Dickhaut, & McCabe, 1995). The trust game measures the degree to which people are willing to invest resources in a trustee partner (here, a human vs. a robot) that is capable of either assisting the investor in attaining rewards or exploiting the investor. Due to its frequent use in labs across the world, we have prior knowledge of how much people normally trust one another and respond to trust, and how much this varies. Previous work has suggested that people make trust game investments in others because of both altruism (i.e., "other-regard") and anticipation of positive reciprocity (i.e., "self-regard") (Cox, 2004). Interactions with autonomous robots can bring about the latter but not the former, because there is no actual human that one might benefit. This other-regarding perspective suggests less investment in robots than in humans (or in robots that affect other humans). However, another possibility is that people make trust-game investments to gain both money and information about humans' behavioral propensities: to learn about the trustworthiness of others (informing how successfully they can develop trust-based exchange relationships) and in anticipation of positive reciprocity (Schniter, Sheremeta, & Shields, 2015). This perspective suggests no investment difference across human or robot trustee, since in both interactions with other humans and in interactions with our robots (programmed to mimic human behavior), participants can also learn about the cooperativeness of fellow humans.

The second way we develop knowledge about trust in robots is by comparing the patterns of emotional reactions generated in trust-based interactions with robots and fellow humans. Below we explain how emotions regulate trust behavior and are triggered by trust interaction outcomes characteristic of distinct adaptive problems (Schniter et al., 2015) – problems which are not equally present across our experimental conditions.

Theory and Predictions

Solving some of the adaptive problems that our human ancestors faced would have required the orchestration of motivation, behavior, physiology, and communication. Emotions appear to be evolved adaptations engineered to orchestrate those systems (Nesse, 1990; Tooby, 1985; Tooby & Cosmides, 2008). An emotion calibrates multiple functionally specialized systems, or adaptations, to those parameters that would have constituted a beneficial response to a recurrent adaptive problem.

Non-social emotions (e.g. contentment, frustration) guide behavior to maximize benefits and minimize costs in the non-social realm. By contrast, social emotions (e.g. gratitude, anger, pride, guilt) respond to successes and failures in the social domain. As we trade off the possible benefits from selfish and cooperative pursuits, we need to integrate the successes and failures of past experiences. Emotions help guide our behavior in decision dilemmas: for example, by recalibrating how much we value another's welfare (Al-Shawaf, Conroy-Beam, Asao, & Buss, 2016; Gómez-Miñambres & Schniter, 2017; Sznycer, Cosmides, & Tooby, 2017; Tooby & Cosmides, 2008).

Trust interactions are decision dilemmas that require decision makers to choose between competing goals that cannot be simultaneously fulfilled at their maxima. From a non-social and individualist perspective, the trust game provides an opportunity for gaining resources. From a social perspective, the trust game also provides valuable information about the trusting and trustworthy character demonstrated by others. We use information about others' cooperativeness to avoid exploitation and develop reliable trust-based exchange relationships that provide a security against income risks associated with luck (Kaplan, Schniter, Smith, & Wilson, 2012, 2018).

Various social and non-social emotions are relevant in the context of trust games. For example, emotions such as anger, gratitude, and guilt have been shown to guide behavior in human-human trust interactions (Schniter & Sheremeta, 2014; Smith, Pedersen, Forster, McCullough, & Lieberman, 2017; Sznycer, 2019). Further, under anonymous conditions in the laboratory, specific emotions such as anger, gratitude, guilt, pride, contentment and frustration result from trust game interactions (Schniter et al., 2015; Schniter & Shields, 2013) and lawfully predict behavior in subsequent trust games with the same person (Schniter & Sheremeta, 2014). Here we extend past work by studying the pattern of activation of social and non-social emotions in trust games with human and robot partners.

Next, we summarily review what is known about the emotions studied here: anger, gratitude, guilt, pride, frustration, and contentment.

Social Emotions

Gratitude is triggered by indications that a partner values one's welfare more than anticipated. Being extended trust and receiving benefits that are costly to deliver are two potent cues that a partner values one's welfare, and therefore can trigger gratitude. Once activated, gratitude functions to cement the cooperative relationship. To do so, the gratitude system up-regulates the value attached to the partner's welfare. Additionally, the gratitude system may signal to its target that this recalibration has taken place. In turn, grateful revaluation of the other's welfare makes it more likely that the partner will continue to cooperate given the opportunity (Algoe, Haidt, & Gable, 2008; Dunn & Schweitzer, 2005; McCullough, Kilpatrick, Emmons, & Larson, 2001; Tesser, Gatewood, & Driver, 1968).

Anger is triggered by cues that a social partner insufficiently values one's welfare. Examples of such cues include: unwillingness to benefit one's self even at low personal cost, and failure to extend the trust that might start or augment a mutually beneficial relationship (Sell et al., 2017). Once activated, anger functions to incentivize the target of the anger to value the angry person's welfare more. To do so, the anger system deploys various types of bargaining tactics designed to render enhanced aid and assistance the more economical option for the target of the anger; for example: threats of (or actual) withdrawal of assistance, and threats of (or actual) aggressive imposition of costs (Sell, 2011).

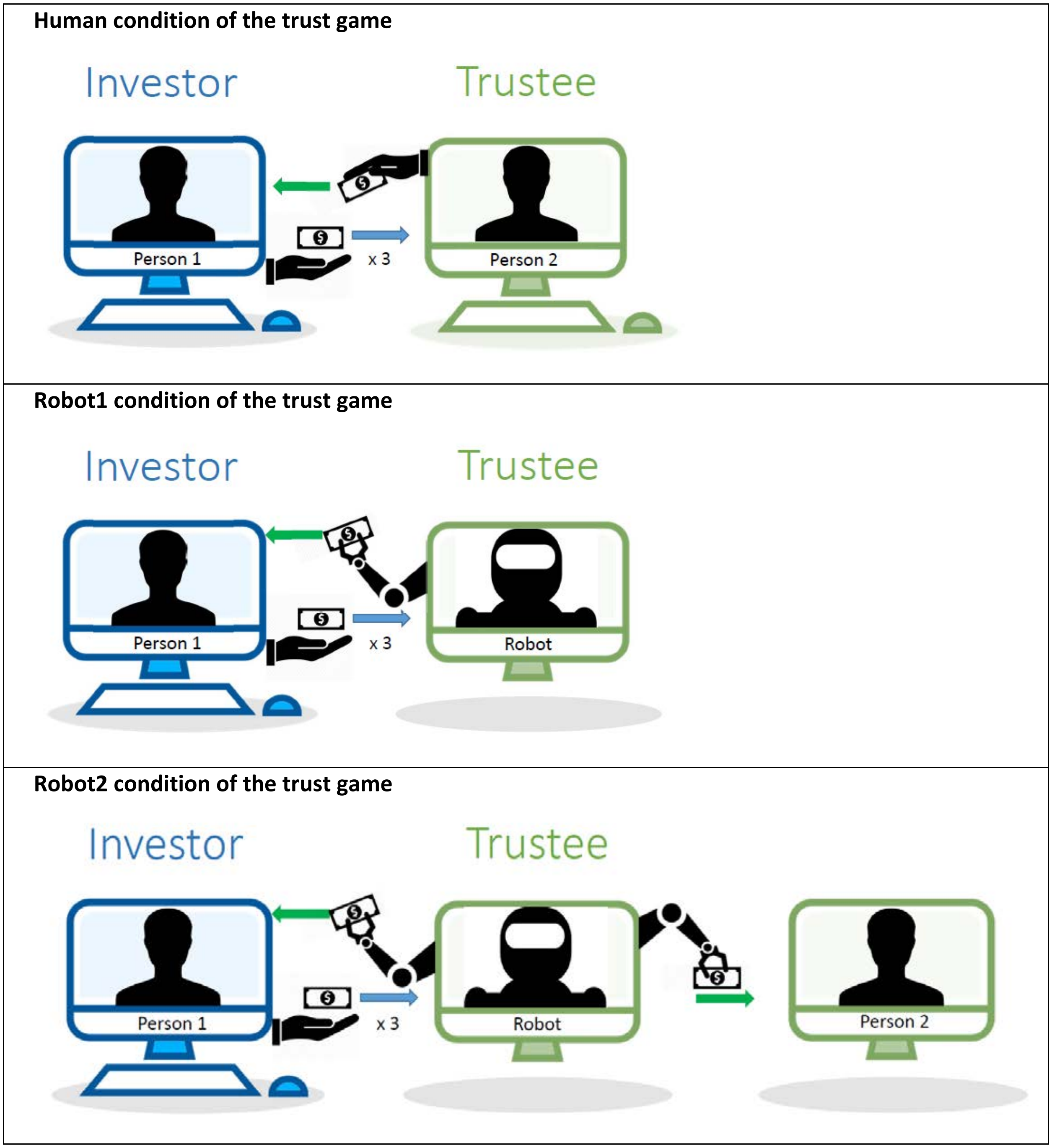
Pride appears to be designed to promote the social value of the individual in the minds of others. Pride motivates the achievement of socially valued acts—acts whose discovery by others would cause them to increase their valuation of the individual's welfare. Once achieved, pride advertises those acts and internally rewards their continuance (Lea & Webley, 1997; Sznycer et al., 2018; Tracy, Shariff, & Cheng, 2010; Williams & DeSteno, 2008). The evolutionary origin of human pride lies in phylogenetically ancient systems undergirding dominance interactions (Fessler, 1999; Weisfeld & Dillon, 2012). Indeed, behavioral dominance is still a reliable source of pride and social status (Cheng, Tracy, Foulsham, Kingstone, & Henrich, 2013). In strategic economic interactions, earning a larger payout can imply economic dominance. Notwithstanding pride's association with dominance, in cooperative contexts the activation of this emotion depends on the individual making positive contributions to others (Cheng et al., 2013; Halevy, Chou, Cohen, & Livingston, 2012; Smith, 1759). Successful cooperation in a trust game provides players the opportunity to benefit themselves by maximizing personal earnings, as well as the opportunity to benefit their partners by investing and reciprocating. Thus, players in a trust game could be proud of the money they earned for themselves and they could be proud because of the benefit they provided their partners.

The mind should positively value another's welfare when the other's welfare is intrinsically valuable to the individual's own welfare (e.g., as is the case with kin, mates, friends) (Tooby & Cosmides, 1996). Insufficiently valuing a valuable other is costly for the offender, even when the under-valued other fails to angrily contest or notice the insufficient valuation (Smith, Webster, Parrott, & Eyre, 2002; Sznycer, 2010; Sznycer et al., 2016). The cost incurred for under-valuing valuable others can be abated by up-regulating one's valuation of their welfare. This is, by hypothesis, the evolved function of guilt. Guilt is triggered by indications that one has insufficiently valued the welfare of a valuable other: for example, by failing to help at relatively low personal cost or by failing to extend trust, for example. When triggered, guilt interrupts and discourages the imposition of costs on the other (Baumeister, Stillwell, & Heatherton, 1995; Cohen, Panter, & Turan, 2013; Cohen, Panter, Turan, Morse, & Kim, 2014; Schniter & Shields, 2013) and motivates actions to benefit the other and to restore the relationship (Baumeister, Stillwell, & Heatherton, 1994; Baumeister et al., 1995): amends, apologies, confessions, and perspective-taking (De Hooge, Zeelenberg, & Breugelmans, 2007; Ketelaar & Au, 2003; Leith & Baumeister, 1998; Ohtsubo & Yagi, 2015; Schniter, Sheremeta, & Sznycer, 2013; Sznycer, Schniter, Tooby, & Cosmides, 2015; Tangney, 1991). For example, trustees who feel guilty about their behavior in previous trust-based interactions are more likely to apologize (Schniter & Sheremeta, 2014).

Non-social emotions

In contrast with the social emotions discussed above, frustration and contentment are triggered by the non-social component of an individual's failures and successes. Frustration is a feeling of dissatisfaction that results when obstacles prevent the achievement of personal goals (Harrington, 2005). Frustration is known to be triggered by interactions with computers that don't result in desired goals (Klein, Moon, & Picard, 2002). Frustration downregulates the motivation to repeat previous behaviors that inhibited the realization of goals (Klein et al., 2002).

Many activities associated with contentment pertain to the individual and are unrelated to social activities (Berenbaum, 2002). Unlike feelings of frustration, which are triggered by things getting in the way of one's goals, feelings of contentment arise when things more successfully go one's way: in situations with greater certainty and a lack of perceived obstacles (Ellsworth & Smith, 1988). Contentment has been identified as a positive feeling of satisfaction, rest, and safety that follows achievement of personal goals (Ellsworth & Smith, 1988; Fredrickson, 1998; Harlé & Sanfey, 2010; Kreibig, 2010). Various lines of research suggest that contentment encourages one to reflect on one's successful goal attainment experiences (Christie & Friedman, 2004). Contentment may also discourage risk taking and other potentially mood changing behaviors (Herzenstein & Gardner, 2009).


Human condition of the trust game

Robot1 condition of the trust game

Robot2 condition of the trust game

Fig. 1. Human, Robot1, and Robot2 conditions of the trust game.

We study trust and emotions across three conditions of the Trust game that we call Human, Robot1, and Robot2. In the Human condition a human participant (Person 1) in the role of investor is paired with a human participant (Person 2) in the role of trustee. In the Robot1 condition a human (Person 1) in the role of investor is paired with a robot in the role of trustee. In the Robot2 condition a human participant (Person 1) in the role of investor is paired with a robot in the role of trustee that acts on behalf of a passive participant (Person 2). Figure 1 depicts the basic experimental set up with conditions, see Methods section for further details.

The goals that non-social emotions such as contentment (or frustration) help guide us towards (or away from) are present in all three conditions of the trust game (Human, Robot1, and Robot2). Higher (or lower) earnings are expect to trigger contentment (or frustration) in participants in all roles. By contrast, social emotions are expected in interactions where human partners are affected directly or indirectly, but not in interactions with robots.

Predictions

We expect no investment difference across conditions (Prediction 1). This is because all conditions provide Person 1s an opportunity to gain both money and information about the trustworthiness of other people in the sample.

What differs across our experimental conditions is how the trustee is *affected* by and *affects* the investor. In interactions with robot trustees, the investor does not interact with a decision maker who experiences emotions and that intentionally decides whether and how much to reciprocate. Furthermore, the robot trustees in this study do not derive any benefits from monetary transfers to them. Beneficiaries of trust-based investment only exist in the Human and Robot2 conditions.

We assess two non-social emotions (contentment and frustration) and four social emotions (guilt, gratitude, anger, and pride) that participants report after trust-based interactions in these Human, Robot1, and Robot2 conditions. We expect no difference across conditions for the experience of non-social emotions: higher personal *earnings* should cause people *contentment*, while lower personal *earnings* should cause their *frustration* (Prediction 2). On the other hand, the social emotions experienced in response to trust games with humans, robots, and robots that affect other humans are expected to differ. Humans recognize robots as unable to understand or respond to social emotions (Gray, Gray, & Wegner, 2007). As a result, humans respond to these emotionally vacuous and automatic agents with indifference to their welfare (Haslam, 2006). For this reason, we expect that interactions with robots will not elicit the same intensity of social emotions that interactions with humans do.

In trust games, a Person 1's decision to invest smaller (or larger) portions of their endowment will directly affect a Person 2's economic opportunity. Because of this, we expect Person 1's investment to positively predict Person 1's pride and negatively predict Person 1's guilt. However, this should be so only in the conditions where a human Person 2 exists (i.e., Human and Robot2 conditions). In contrast, these relationships should be weaker or null in the Robot1 condition, where no human Person 2 exists (Prediction 3). We expect the Person 2 (the trustee in the Human condition and the passive beneficiary in the Robot2 condition) to react emotionally with *gratitude* (or *anger*) in response to Person 1's higher (or lower) *investment* (Prediction 4).

When *reciprocation* occurs, Person 1s make gains from investments in trustees. When *non-reciprocation* occurs, Person 1s suffer losses on investments in trustees. Being the target of *non-reciprocation* causes Person 1's *anger*, while being the beneficiary of *reciprocation* causes Person 1s *gratitude* in the Human condition. However, because Person 1 is not expected to hold the passive human beneficiary in the Robot2 condition responsible for the automated reciprocity decision made by the robot, this effect should not be as strong in the Robot1 and Robot2 conditions (Prediction 5). From the perspective of a Person 2 who has made a *reciprocation* (or *non-reciprocation*) decision in the role of trustee, pride (or guilt) is expected. In the Robot2 condition, we do not expect these *pride* or *guilt* effects to be as strong; the Person 2s are passive (making no decisions) and the *reciprocation* or *non-reciprocation* decision affecting both players is automated (Prediction 6). We provide a summary of these predictions in Table 1.

Table 1. Predictions for Human, Robot1, and Robot2 conditions of the trust game.

Prediction 1	**No difference in *investment* is expected across conditions.**
Prediction 2	*Contentment* (*frustration)* is positively correlated with higher(lower) personal *earnings* for Person 1 and Person 2 in all conditions.
Prediction 3	Higher (or lower) *investment* affects Person 1 *pride* (*guilt*) more in the Human and Robot2 condition.
Prediction 4	Higher (or lower) *investment* affects Person 2 *gratitude* (*anger*) more in the Human and Robot2 condition.
Prediction 5	*Reciprocation* (*non-reciprocation*) affects Person 1 *gratitude* (*anger*) more in the Human condition.
Prediction 6	*Reciprocation* (*non-reciprocation*) affects Person 2 *pride* (*guilt*) more in the Human condition.

Material and methods

We recruited a convenience sample of 387 participants from a campus-wide subject pool consisting primarily of undergraduate students. Participants who previously participated in trust experiments were not recruited. Using a between-subjects design, we compared behaviors and emotions from human investors (n = 85) in a trust game with human trustees (n=85) (the "Human" condition) to behaviors and emotions from two nearly identical trust games that we call the Robot1 and Robot2 conditions. The Robot1 trust game features human investors (n = 71) and a virtual trustee that we call the "robot". The Robot2 trust game features human investors (n = 73) and robot trustees linked to a passive human beneficiary of the robot's payoff that otherwise does not decide or do anything (n = 73).

The experiment was programmed using z-Tree (Fischbacher, 2007). No participant participated more than once. Each session had between 10 and 24 participants, seated in individual cubicles, and lasted approximately 35 minutes. Sessions were conducted as follows. An experimenter read the instructions aloud explaining experimental procedures and payoffs while participants followed along with their own copy of the instructions. After finishing the instructions, participants were given five minutes to privately answer a quiz testing their comprehension of the instructions. After participants completed the quiz, the experimenter distributed a printed copy of the correct answers. To ensure understanding, the experimenter answered any remaining questions privately.

Participants, randomly assigned to the roles of ''Person 1" (investor) or ''Person 2" (the trustee in the "Human" condition, or the passive recipient in the "Robot2" condition), interacted anonymously in the trust game over a local computer network. Earnings from the trust game plus a $7 participation fee were paid out privately at the end of the experiment.

In the baseline trust game ("Human" condition), participants were randomly assigned to roles of "Person 1" or "Person 2". Person 1 was endowed with $10 and could transfer any portion to Person 2, which was tripled upon receipt. Person 2 then chose to return to Person 1 a portion of the amount received. In our trust games with robots, all participants in the Person 1 role had the option to send any whole dollar portion of their endowment to a "robot", with the amount sent tripled. Participants in the Robot1 condition were told that after they transfer an amount *x,* the "robot" randomly selects a return amount to transfer back to Person 1 from those in a set of observed returns made by people in response to tripled transfers of amount *x* in a previous trust game experiment (the Human condition; see instructions Online Supplementary Materials A). In the Robot2 condition, participants were assigned to roles of Person 1 or

Person 2, with the Person 2 role different than in the Human condition. In the Robot2 condition we explained the role played by the robot to all participants, as above, but with the additional feature that any of the tripled transfer amount not returned to Person 1 by the robot would be kept by a passive Person 2 who makes no decisions.

After playing the trust game, participants completed a 20-item emotion survey (see Online Supplementary Materials B) based on the Positive and Negative Affect Schedule (PANAS) (Watson, Clark, & Tellegen, 1988). Participants used a 5-point Likert scale to report the extent to which they felt each of various emotions.

Results

Consistent with prior results in the literature, we observed substantial variability in individual behavior, with many participants showing trust with investment and trustworthiness with reciprocation. **Figure 2** displays the bubble plot of the amount invested and the amount returned for 387 participants in the Human, Robot1, and Robot2 conditions. Overall, our results are consistent with previous findings of Berg et al. (1995) and reported in panel A of Table 2. On average, Person 1 invested and was returned comparable amounts across conditions, resulting in comparable earnings for Person 1 and Person 2 across conditions. We find no significant differences in Person 1 earnings (p = .68) or Person 2 earnings (p = .93) across conditions using the Kruskal-Wallis equality-of-populations rank test (hereafter Kruskal-Wallis). For interactions with investments greater than zero we find the effects of net "Reciprocation" (return-investment) and net "Non-reciprocation" (investment-return) to be comparable across conditions using Kruskal-Wallis (p = .64).

Comparing participants' interactions with humans to interactions with robots, we highlight important similarities: humans and robots are trusted similarly across conditions (Result A) and non-social emotions are similarly triggered by one's own earnings (Result B). We also highlight important differences in social emotions resulting from Human, Robot1, and Robot2 trust game interactions (Result C).

Investment

Result A: Investment amounts by Person 1 are similar across the Human, Robot1, and Robot2 conditions (Prediction 1).

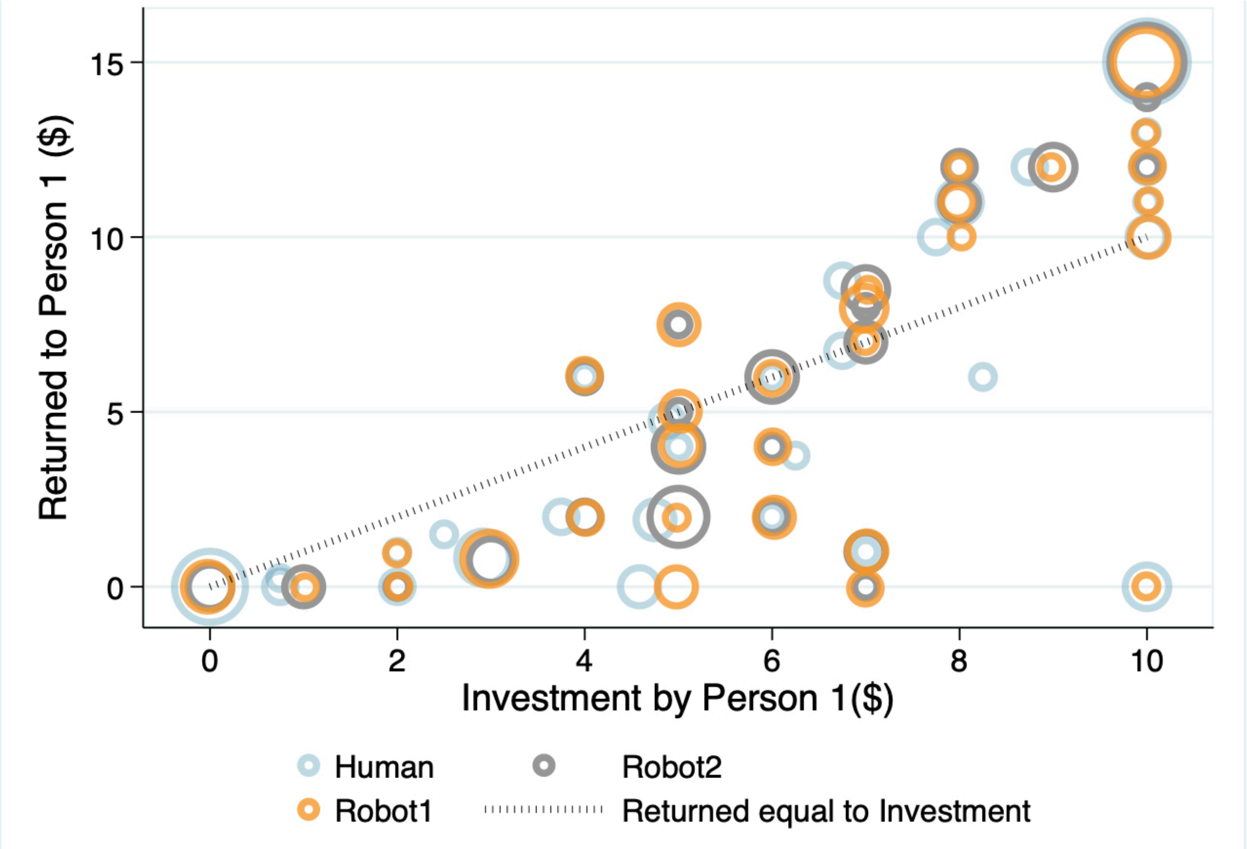


Fig. 2. Bubble plot of the amount invested and the amount returned by trustee to Person 1, by condition. Note: Relative size of bubbles indicate the number of observations. Smallest bubble = 1 observation; largest bubble = 8 observations.

Similar distributions of investment amounts are seen across conditions. A Kruskal-Wallis test indicates no significant differences in investment distributions between conditions (*p* = .99). Provided a novel trust-based exchange opportunity, participants are as likely to trust humans (N = 85, Mdn = $6.5), as they are to trust robots in the Robot1 condition (N = 71, Mdn investment = $6.0) or Robot2 condition (N = 73, Mdn = $6.0); a Mood's median test finds no differences in median investment across conditions (*p* = .183).

In addition, we find no significant differences in the fraction of participants who choose not to invest across conditions (13%, 7% and 5% for the Human, Robot1, and Robot2 conditions, respectively) using a Kruskal-Wallis test (*p* = .212). Likewise, we find no significant differences in the fraction of participants who choose to investment the maximum across conditions (32%, 24% and 19% for the Human, Robot1, and Robot2 conditions, respectively) using a Kruskal-Wallis test (*p* = .184).

Table 2.

Panel A: Independent Variables used in OLS Regressions (Mean and *SD* reported)

		Earnings		Reciprocation/
Treatment	Investment	Person 1	Person 2	Non-reciprocation*
Human	6.01	10.14	11.88	0.17
N = 85 (74*)	*3.64*	*3.72*	*7.12*	*3.98*
Robot1	6.11	9.79		-0.23
N = 71 (66*)	*2.99*	*3.4*		*3.53*
Robot2	6.18	10.4	11.95	0.43
N = 73 (69*)	*2.86*	*3.17*	*4.64*	*3.26*
Kruskal-Wallis	$\chi^2(2) = 0.01$	$\chi^2(2) = 0.76$	$\chi^2(1) < 0.01$	$\chi^2(2) = 0.91$
rank test	*p* = .999	*p* = .684	*p* = .932	*p* = .635

Panel B: Dependent Variables used in OLS Regressions (Mean and *SD* reported)

	Contentment		Frustration		Gratitude		Anger		Pride		Guilt	
	Person	Person	Person	Person	Person	Person	Person	Person	Person	Person	Person	Person
Treatment	1	2	1	2	1*	2	1*	2	1	2*	1	2*
Human	3.44	3.13	1.74	2.01	3.07	3.21	1.72	1.88	2.76	2.93	1.53	1.86
N = 85 (74*)	*1.35*	*1.45*	*1.26*	*1.48*	*1.64*	*1.63*	*1.16*	*1.34*	*1.33*	*1.56*	*1.02*	*1.34*
Robot1	3.25		2.14		2.80		1.80		2.24		1.34	
N = 71 (66*)	*1.32*		*1.29*		*1.36*		*1.10*		*1.20*		*0.75*	
Robot2	3.26	3.30	1.73	1.74	3.17	3.27	1.51	1.53	2.41	2.22	1.40	1.22
N = 73 (69*)	*1.19*	*1.20*	*1.07*	*1.16*	*1.29*	*1.39*	*0.82*	*1.07*	*1.36*	*1.35*	*0.95*	*0.72*

NOTE: Asterisked columns exclude observations where Person 2 received zero so could not take an action, resulting in a smaller number of observations. Reciprocation is the amount returned to Person 1 minus the amount invested by Person 1. Non-reciprocation is equal to Reciprocation with the sign reversed.

Non-social emotions

Result B: Non-social emotions (*contentment, frustration*) are similarly triggered by one's own earnings across the Human, Robot1, and Robot2 conditions (Prediction 2).

Contentment resulting from greater personal *earnings* was predicted to be similar for Person 1 and for Person 2 across conditions—because the non-social emotional response to personal gains should depend on the intrinsic properties of those gains (here, their magnitude), equally present in all conditions, and not on whether those gains were caused by a human partner, an algorithm, or some other cause that varies by condition. This is what we observed. Person 1 and Person 2 reported feeling more *content* the higher their personal *earnings* were from the interaction (see **Figure 3 panels A and B**). We report the mean and standard deviation of independent and dependent variables in Table 2, and the results of OLS regression, including standardized beta coefficients, below.

For Person 1, contentment increased with earnings in the Human condition (β_{Human} = .577, t(84) = 6.44, p < .001), the Robot1 condition (β_{Robot1} = .281, t(70) = 2.44, p = .017), and the Robot2 condition (β_{Robot2} = .393, t(72) = 3.60, p = .001). Moreover, the relationship between Person 1's *contentment* and their *earnings* did not differ significantly across conditions. We compared the coefficients of Person 1's *contentment* as a function of their *earnings* across conditions to test the null hypotheses Ho$_1$: $\beta_{Human} = \beta_{Robot1}$ and Ho$_2$: $\beta_{Human} = \beta_{Robot2}$. We find a marginal difference between Human and Robot1 conditions (t(155) = -1.900, p = .059), and no difference between Human and Robot2 conditions (t(157) = -1.130, p = .261).

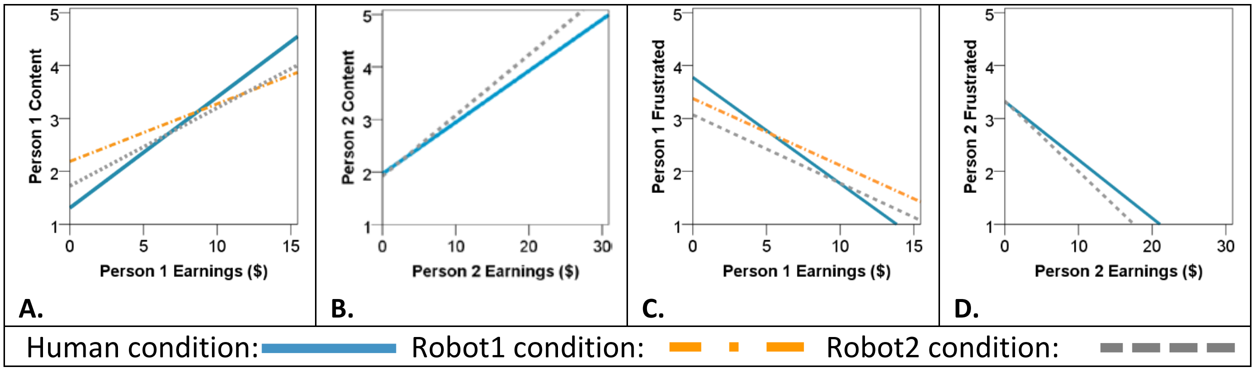


Fig. 3. Non-social emotions as a function of earnings

For Person 2, positive relationships between their *earnings* and *contentment* were observed in the Human condition (β_{Human} =.482, t(84) = 5.01, p < .001) and in the Robot2 condition (β_{Robot2} =.446, t(72) = 4.20, p <

.001). We compared the coefficients of Person 2's *contentment* as a function of Person 2's *earnings* across conditions to test the null hypothesis Ho$_1$: β$_{Human}$ = β$_{Robot2}$. We find no difference between Human and Robot2 conditions (*t*(157) = 0.49, *p* = .622).

Frustration resulting from smaller personal *earnings* was predicted to be similar for Person 1 and for Person 2 across conditions. This is what we observed. Person 1 and Person 2 reported feeling more *frustrated* the lower their personal *earnings* were from the interaction (see **Figure 3 panels C and D**).

For Person 1, frustration decreased with earnings in the Human condition (β$_{Human}$ = -.594, t(84) = -6.72, *p* < .001), the Robot1 condition (β$_{Robot1}$ = -.333, t(70) = -2.93, *p* = .005), and the Robot2 condition (β$_{Robot2}$ = -.382, t(72) = -3.49, *p* = .001). Moreover, the relationship between Person 1's *frustration* and their lower *earnings* did not differ significantly across conditions. We compared the coefficients of Person 1 *frustration* as a function of their lower *earnings* across conditions to test the null hypotheses Ho$_1$: β$_{Human}$ = β$_{Robot1}$ and Ho$_2$: β$_{Human}$ = β$_{Robot2}$. We find no difference between Human and Robot1 conditions (*t*(155) = 1.51, *p* = .133), and no difference between Human and Robot2 conditions (*t*(157) = 1.40, *p* = .163).

For Person 2, the negative relationship between their *earnings* and *frustration* was observed both in the Human condition (β$_{Human}$ = -.529, t(84)= -5.68, *p* < .001), and in the Robot2 condition (β$_{Robot2}$ = -.534, t(72)= -5.32, *p* < .001). We compared the coefficients of Person 2's *frustration* as a function of Person 2's *earnings* across conditions to test the null hypothesis Ho$_1$: β$_{Human}$ = β$_{Robot2}$. We find no difference between Human and Robot2 conditions (*t*(157) =.67, *p* =.505).

Social emotions

Result C: Social emotions (*gratitude, anger, pride, guilt*) resulting from trust game interactions sometimes differed across the Human, Robot1, and Robot2 conditions (Predictions 3-6).

Higher *investment* predicted Person 1's *pride* to a greater extent in the Human and Robot2 conditions than in the Robot1 condition—because in the former two conditions *investment* provides a human partner the opportunity to benefit, whereas in the latter condition there is no human partner (Prediction 3). This is what we observed (see **Figure 4 Panel A**). Person 1 reported feeling more *pride* with higher *investments* in the Human (β$_{Human}$ = .345, t(84) = 3.35, *p* = .001) and Robot2 conditions (β$_{Robot2}$ = .355, t(72) = 3.20, *p* = .002); but this relationship was not significant in the Robot1 condition (β$_{Robot1}$ = .159, t(70) = 1.34, *p* = .184). Higher Person 1 *earnings* predicted Person 1's *pride* in all conditions equally because they each provide equal opportunity for personal gain. We observed this as well. There is support for Person 1 reporting more *pride* with higher personal *earnings* in the Human (β_{Human} = .324, t(84) = 3.12, *p* = .002), Robot1 (β_{Robot1} = .493, t(70) = 4.71, *p* < .001), and Robot2 conditions (β_{Robot2} = .467, t(72)=4.45, *p* < .001).

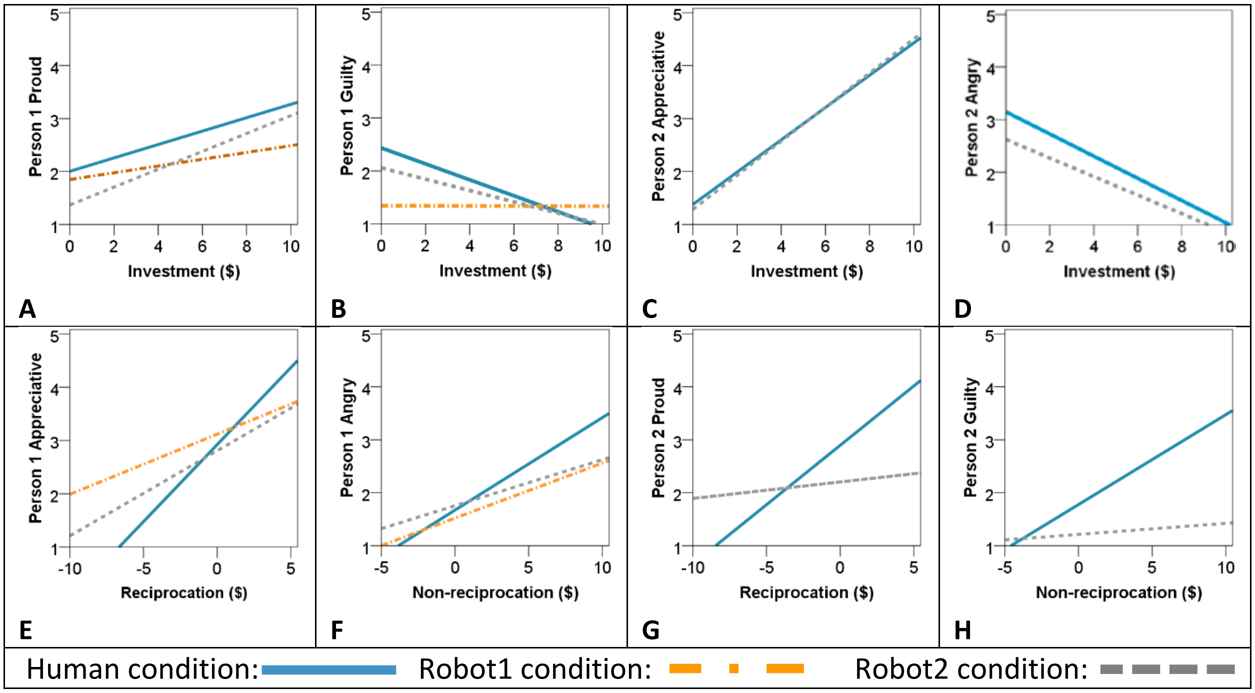


Fig. 4. Social emotions as a function of investment, reciprocation (= return - investment), and non-reciprocation (= investment - return), by condition.

Lower *investment* predicted Person 1s *guilt* to a greater extent in the Human and Robot2 conditions than in the Robot1 condition—because in the former two conditions failure to invest deprives a human partner from the opportunity to benefit, whereas in the latter condition there is no human partner (Prediction 3). This is what we observed (see **Figure 4 Panel B**). Person 1 reported feeling less *guilty* with higher *investments*. This was true in the Human condition (β_{Human} = -.538, t(84) = -5.81, *p* < .001) and in the Robot2 condition (β_{Robot2} = -.321, t(72) = -2.86, *p* = .006), but not in the Robot1 condition (β_{Robot1} = -.004, t(70) = -0.04, *p* = .971). Moreover, the relationship between Person 1's *guilt* and their level of *investment* differed significantly between the Human condition and the Robot2 condition on the one hand, and the Robot1 condition on the other hand (*t*(228) = 3.40, *p* = .001).

Higher *investment* predicted Person 2's *gratitude* in both the Human and Robot2 conditions—because Person 1's investment provides Person 2 the opportunity to benefit (Prediction 4). This is what we observed (see **Figure 4 Panel C**). Person 2s reported feeling more *gratitude* with higher *investment* both in the Human condition (β_{Human} = .679, t(84) = 8.43, *p* < .001) and in the Robot2 condition (β_{Robot2} = .666, t(72) = 7.52, *p* < .001). Moreover, the relationship between Person 2's *gratitude* and higher *investment* from Person 1 did not differ significantly between the Human condition and the Robot2 condition ($t(157) = 0.30$, $p = .764$).

Lower *investment* predicted Person 2's *anger* in the Human and Robot2 conditions because in these conditions Person 1's investment provides a human partner the opportunity to benefit (Prediction 4). This is what we observed (see **Figure 4 Panel D**). Person 2s reported feeling more *angry* with lower *investments* in both the Human condition ($\beta_{\text{Human}} = -.571$, t(84) = -6.33, $p < .001$) and Robot2 condition ($\beta_{\text{Robot2}} = -.472$, t(72) = - 4.51, $p < .001$). Further, the relationship between Person 2's *anger* and lower *investment* was no different across the Human and Robot2 conditions ($t(157) = 0.65$, $p = .516$).

Higher *reciprocation* predicted Person 1's *gratitude* in the Human condition (where a partnered Person 2's decision positively affected Person 1's gains from investment), but not in the Robot1 and Robot2 conditions (where reciprocation decisions were automated) (Prediction 5). Considering only observations where Person 2 could make a reciprocation decision (cases where Person 1 invested more than $0), this is what we observed (see **Figure 4 Panel E**). Person 1s reported feeling more *gratitude* when *reciprocation* from the other party was higher. These results were stronger in the Human condition ($\beta_{\text{Human}} = .699$, t(73) = 8.29, $p < .001$) than in the Robot1 condition ($\beta_{\text{Robot1}} = .416$, t(65) = 3.66, $p = .001$) or the Robot2 condition ($\beta_{\text{Robot2}} =.284$, t(68) = 2.42, $p = .018$). Further, the relationship between Person 1's *gratitude* and higher *reciprocation* differed significantly between the Human condition on the one hand, and the Robot1 and Robot2 conditions on the other hand ($t(208) = 3.07$, $p = .002$).

Higher *non-reciprocation* predicted Person 1's *anger* where a partnered Person 2's decision affected Person 1's losses from investment, but not in the Robot1 and Robot2 conditions where reciprocation decisions were automated (Prediction 5). Considering only observations where Person 2 could make a decision (cases where Person 1 invested more than $0), this is what we observed (see **Figure 4 Panel F**). Person 1s reported feeling more *angry* with higher *non-reciprocation*. This relationship was stronger in the Human condition ($\beta_{\text{Human}} = .600$, t(73) = 6.36, $p < .001$) than in the Robot1 condition ($\beta_{\text{Robot1}} = .277$, t(65) = 2.31, $p = .024$) or the Robot2 condition ($\beta_{\text{Robot2}} = .420$, t(68) = 3.79, $p < .001$). And, importantly, the relationship between Person 1's *anger* and being the victim of *non-reciprocation* differed significantly between the Human condition on the one hand, and the Robot1 and Robot2 conditions on the other hand (t(208) = 2.144, $p = .033$).

Higher *reciprocation* predicted Person 2's *pride* to a greater extent in the Human condition than in the Robot2 condition (Prediction 6). In the Human condition Person 2 decides to reciprocate determining

Person 1's gains from investments, but in the Robot2 condition Person 2 makes no such decision. Having dropped observations when Person 2 received nothing and only considering only observations where Person 2 could make a reciprocation decision, this is what we observed (see **Figure 4 Panel G**). Person 2 reported feeling more *pride* with higher *reciprocation* in the Human condition (β_{Human} = .571, t(73) = 5.90, *p* < .001) but not in the Robot2 condition (β_{Robot2} = .076, t(68) = 0.62, *p* = .537). Higher Person 2 *earnings* also predicted Person 2's *pride* in both Human and Robot2 conditions equally because they each provide equal opportunity for personal gain. This is what we observed. There is support for Person 2 reporting more *pride* with higher personal *earnings* in the Human (β_{Human} = .282, t(84) = 2.678, *p* = .009) and Robot2 conditions (β_{Robot2} = .428, t(72) = 3.99, *p* < .001). And, critically, the relationship between Person 2's *pride* and higher *reciprocation* differed significantly between the Human condition and the Robot2 condition (*t*(143) = 3.08, *p* = .003) despite the strong main effect of personal *earnings* on pride in both conditions.

Higher *non-reciprocation* predicted Person 2's *guilt* to a greater extent in the Human condition than in the Robot2 condition (Prediction 6). In the Human condition, Person 2's non-reciprocation decision causes Person 1's loss on investment, but in the Robot2 condition Person 2 makes no such decision. We drop observations when Person 2 received nothing, and, considering only observations where Person 2 could make a decision, this is what we observed (see **Figure 4 Panel H**). Person 2 reported feeling more *guilty* with higher *non-reciprocation* in the Human condition (β_{Human} = .509, t(73) = 5.02, *p* < .001) but not in the Robot2 condition (β_{Robot2} = .096, t(68) = 0.79, *p* = .433). And, critically, the relationship between Person 2's *guilt* and *reciprocation* to Person 1 differed significantly between the Human condition and the Robot2 condition (*t*(142) = 3.24, *p* = .002).

Discussion

Natural selection designed emotions to solve adaptive problems by recalibrating cognition and behavior in fitness-promoting ways (Cosmides & Tooby, 2000). As we increasingly put our trust in robots it is important to better understand our emotional reactions to interactions with them. Our study demonstrates yet another example of humans willing to take risks to engage in trust-based interactions with robots– similar to how they do with one another—yet experiencing different emotional reactions to those interactions. These differences were as predicted.

Future research should examine how social emotions regulate trust conditional on past interactions with one's partner, and whether trust-re-extension varies according to robot type. Above we report that trust-based interactions with fellow humans and with robots affecting fellow humans elicit more intense social emotions—positive (pride and gratitude) and negative (guilt and anger)—compared to trust-based interactions with robots acting alone.

Given that investors' initial trust did not differ across trustee type but investors' social emotions did, a distinct possibility is that trust re-extension will differ when the trustee is a human, a robot, or a robot linked to other human beneficiaries. For example, partnerships with consistent reciprocators may consolidate into stronger, more productive partnerships when the reciprocators are fellow humans, because humans elicit more gratitude than robots do. Conversely, partnerships with inconsistent reciprocators may be more stable when the reciprocators are robots, because robots elicit less anger than humans do. Further, humans experienced pride and guilt more intensely in interactions where robots served a beneficiary than where robots acted alone. This suggests that people will be even more likely to re-extend trust to robot partners when those robots link to other human beneficiaries.

The human cognitive architecture evolved to have enough structure and content to promote our ancestors' survival and reproduction while also having the flexibility to navigate novel challenges and opportunities. These features enable humans to design and rationally interact with AI and robots—agents whom our forager ancestors may never have imagined. Still, interactions with automata, and science's ability to explain these interactions, are imperfect, because automata (i) lack the psychophysical cues that we evolved to expect in an interaction partners and (ii) often are guided by unintuitive decision logics. Our abilities to design, build, and interact with automata are testament to the power of human cognition. A behavioral science of human social interactions with AI and robots that harnesses this power has great promise.

REFERENCES

Algoe, S., Haidt, J., & Gable, S. (2008). Beyond Reciprocity. *Emotion*, *8*(3), 425–429. https://doi.org/10.1037/1528-3542.8.3.425

Al-Shawaf, L., Conroy-Beam, D., Asao, K., & Buss, D. M. (2016). Human Emotions: An Evolutionary Psychological Perspective. *Emotion Review*, *8*(2), 173–186. https://doi.org/10.1177/1754073914565518

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243–267. https://doi.org/10.1037/0033-2909.115.2.243

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1995). Personal narratives about guilt: Role in action control and interpersonal relationships. *Basic and Applied Social Psychology*, *17*(1–2), 173–198.

Berenbaum, H. (2002). Varieties of joy-related pleasurable activities and feelings. *Cognition and Emotion*, *16*(4), 473–494. https://doi.org/10.1080/0269993014000383

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology*, *104*(1), 103.

Christie, I. C., & Friedman, B. H. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. *International Journal of Psychophysiology*, *51*(2), 143–153.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence by Andy Clark*. New York: Oxford University Press.

Cohen, T. R., Panter, A. T., & Turan, N. (2013). Predicting Counterproductive Work Behavior from Guilt Proneness. *Journal of Business Ethics*, *114*(1), 45–53. https://doi.org/10.1007/s10551-012-1326-2

Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2014). Moral character in the workplace. *Journal of Personality and Social Psychology*, *107*(5), 943.

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In *Handbook of Emotions. M. Lewis & J. M. Haviland-Jones (Eds.)* (2nd ed., pp. 91–115). New York: Guilford.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281.

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion*, *21*(5), 1025–1042.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and Believing: The Influence of Emotion on Trust. *Journal of Personality and Social Psychology*, *88*(5), 736–748. https://doi.org/10.1037/0022-3514.88.5.736

Ellsworth, P. C., & Smith, C. A. (1988). Shades of Joy: Patterns of Appraisal Differentiating Pleasant Emotions. *Cognition and Emotion*, *2*(4), 301–331. https://doi.org/10.1080/02699938808412702

Fessler, D. M. T. (1999). The Universality of Second Order Emotions. In *Beyond Nature or Nurture: Biocultural Approaches to the Emotions. Hinton, Alexander Laban (ed.)* (pp. 75–116). New York: Cambridge University Press.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

Fredrickson, B. L. (1998). What Good Are Positive Emotions? *Review of General Psychology*, *2*(3), 300–319. https://doi.org/10.1037/1089-2680.2.3.300

Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: Experiments in e-Products and e-Services. *Omega*, *32*(6), 407–424. https://doi.org/10.1016/j.omega.2004.01.006

Gómez-Miñambres, J., & Schniter, E. (2017). Emotions and Behavior Regulation in Decision Dilemmas. *Games*, *8*(2), 22. https://doi.org/10.3390/g8020022

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies*, *8*(3), 483–500. https://doi.org/10.1075/is.8.3.10gro

Halevy, N., Chou, E. Y., Cohen, T. R., & Livingston, R. W. (2012). Status conferral in intergroup social dilemmas: Behavioral antecedents and consequences of prestige and dominance. *Journal of Personality and Social Psychology*, *102*(2), 351–366. https://doi.org/10.1037/a0025515

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Harlé, K. M., & Sanfey, A. G. (2010). Effects of approach and withdrawal motivation on interactive economic decisions. *Cognition and Emotion*, *24*(8), 1456–1465. https://doi.org/10.1080/02699930903510220

Harrington, N. (2005). The Frustration Discomfort Scale: Development and psychometric properties. *Clinical Psychology & Psychotherapy*, *12*(5), 374–387. https://doi.org/10.1002/cpp.465

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

Herzenstein, M., & Gardner, M. (2009). All Positive Emotions Are Not Created Equal: The Case of Joy and Contentment. *ACR North American Advances*, *NA-36*. Retrieved from http://acrwebsite.org/volumes/14311/volumes/v36/NA-36

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889. https://doi.org/10.1016/j.joep.2011.05.007

Kaplan, H. S., Schniter, E., Smith, V. L., & Wilson, B. J. (2012). Risk and the evolution of human exchange. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1740), 2930–2935. https://doi.org/10.1098/rspb.2011.2614

Kaplan, H. S., Schniter, E., Smith, V. L., & Wilson, B. J. (2018). Experimental tests of the tolerated theft and risk-reduction theories of resource exchange. *Nature Human Behaviour*, *2*(6), 383. https://doi.org/10.1038/s41562-018-0356-x

Ketelaar, T., & Au, W. T. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, *17*(3), 429–453. https://doi.org/10.1080/02699930143000662

Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, *14*(2), 119–140. https://doi.org/10.1016/S0953-5438(01)00053-4

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Lea, S. E. G., & Webley, P. (1997). Pride in economic psychology. *Journal of Economic Psychology*, *18*(2), 323–340. https://doi.org/10.1016/S0167-4870(97)00011-1

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, J. J., Knox, B., Baumann, J., Breazeal, C., & DeSteno, D. (2013). Computationally Modeling Interpersonal Trust. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00893

Leith, K. P., & Baumeister, R. F. (1998). Empathy, Shame, Guilt, and Narratives of Interpersonal Conflicts: Guilt-Prone People Are Better at Perspective Taking. *Journal of Personality*, *66*(1), 1–37. https://doi.org/10.1111/1467-6494.00001

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

McCullough, M. E., Kilpatrick, S. D., Emmons, R. A., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin*, *127*(2), 249–266. https://doi.org/10.1037/0033-2909.127.2.249

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature*, *1*(3), 261–289. https://doi.org/10.1007/BF02733986

Ohtsubo, Y., & Yagi, A. (2015). Relationship value promotes costly apology-making: Testing the valuable relationships hypothesis from the perpetrator's perspective. *Evolution and Human Behavior*, *36*(3), 232–239. https://doi.org/10.1016/j.evolhumbehav.2014.11.008

Rai, T. S., & Diermeier, D. (2015). Corporations are Cyborgs: Organizations elicit anger but not sympathy when they can think but cannot feel. *Organizational Behavior and Human Decision Processes*, *126*, 18–26. https://doi.org/10.1016/j.obhdp.2014.10.001

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of Robots in Emergency Evacuation Scenarios. *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, 101–108. Retrieved from http://dl.acm.org/citation.cfm?id=2906831.2906851

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 141–148. https://doi.org/10.1145/2696454.2696497

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schniter, E., & Sheremeta, R. M. (2014). Predictable and predictive emotions: Explaining cheap signals and trust re-extension. *Frontiers in Behavioral Neuroscience*, *8*. https://doi.org/10.3389/fnbeh.2014.00401

Schniter, E., Sheremeta, R. M., & Shields, T. W. (2015). Conflicted emotions following trust-based interaction. *Journal of Economic Psychology*, *51*, 48–65. https://doi.org/10.1016/j.joep.2015.08.006

Schniter, E., Sheremeta, R. M., & Sznycer, D. (2013). Building and rebuilding trust with promises and apologies. *Journal of Economic Behavior & Organization*, *94*, 242–256. https://doi.org/10.1016/j.jebo.2012.09.011

Schniter, E., & Shields, T. W. (2013). Recalibrational Emotions and the Regulation of Trust-Based Behaviors. In *Psychology of Trust; New Research,*. Rochester, NY: D. Gefen (ed.), Nova Science Publishers.

Sell, A. N. (2011). The recalibrational theory and violent anger. *Aggression and Violent Behavior*, *16*(5), 381–389. https://doi.org/10.1016/j.avb.2011.04.013

Sell, A. N., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., … Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128. https://doi.org/10.1016/j.cognition.2017.06.002

Smith, A. (1759). *The Theory of Moral Sentiments* (2002 printing). Cambridge University Press.

Smith, A., Pedersen, E. J., Forster, D. E., McCullough, M. E., & Lieberman, D. (2017). Cooperation: The roles of interpersonal value and gratitude. *Evolution and Human Behavior*, *38*(6), 695–703. https://doi.org/10.1016/j.evolhumbehav.2017.08.003

Smith, R. H., Webster, J. M., Parrott, W. G., & Eyre, H. L. (2002). The role of public exposure in moral and nonmoral shame and guilt. *Journal of Personality and Social Psychology*, *83*(1), 138–159. https://doi.org/10.1037/0022-3514.83.1.138

Sznycer, D. (2010). *Cognitive adaptations for calibrating welfare tradeoff motivations, with special reference to the emotion of shame.* (University of California, Santa Barbara). Retrieved from https://search.proquest.com/openview/1635a342089cc0a5156875803a039efc/1?pq-origsite=gscholar&cbl=18750&diss=y

Sznycer, D. (2019). Forms and Functions of the Self-Conscious Emotions. *Trends in Cognitive Sciences*, *23*(2), 143–157. https://doi.org/10.1016/j.tics.2018.11.007

Sznycer, D., Cosmides, L., & Tooby, J. (2017). Adaptationism Carves Emotions at Their Functional Joints. *Psychological Inquiry*, *28*(1), 56–62. https://doi.org/10.1080/1047840X.2017.1256132

Sznycer, D., Schniter, E., Tooby, J., & Cosmides, L. (2015). Regulatory adaptations for delivering information: The case of confession. *Evolution and Human Behavior*, *36*(1), 44–51. https://doi.org/10.1016/j.evolhumbehav.2014.08.008

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences*, *113*(10), 2625–2630. https://doi.org/10.1073/pnas.1514699113

Sznycer, D., Xygalatas, D., Alami, S., An, X.-F., Ananyeva, K. I., Fukushima, S., … Tooby, J. (2018). Invariances in the architecture of pride across small-scale societies. *Proceedings of the National Academy of Sciences*, *115*(33), 8322–8327. https://doi.org/10.1073/pnas.1808418115

Tangney, J. P. (1991). Moral affect: The good, the bad, and the ugly. *Journal of Personality and Social Psychology*, *61*(4), 598–607. https://doi.org/10.1037/0022-3514.61.4.598

Tesser, A., Gatewood, R., & Driver, M. (1968). Some determinants of gratitude. *Journal of Personality and Social Psychology*, *9*(3), 233–236. https://doi.org/10.1037/h0025905

Tooby, J. (1985). The Emergence of Evolutionary Psychology. In *Emerging Syntheses In Science, David Pine (ed.)* (pp. 67–76). https://doi.org/10.1201/9780429492594-6

Tooby, J., & Cosmides, L. (1996). Friendship and the Banker's Paradox: Other pathways to the evolution of adaptations for altruism. In *Proceedings of the British Academy*: *Vol. 88*. *Evolution of Social Behaviour Patterns in Primates and Man. W. G. Runciman, J. Maynard Smith, & R. I. M. Dunbar (Eds.),* (pp. 119–143).

Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *Handbook of emotions, 3rd ed* (pp. 114–137). New York, NY, US: The Guilford Press.

Tracy, J. L., Shariff, A. F., & Cheng, J. T. (2010). A Naturalist's View of Pride. *Emotion Review*, *2*(2), 163–177. https://doi.org/10.1177/1754073909354627

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037//0022-3514.54.6.1063

Weisfeld, G. E., & Dillon, L. M. (2012). Applying the dominance hierarchy model to pride and shame, and related behaviors. *Journal of Evolutionary Psychology*, *10*(1), 15–41. https://doi.org/10.1556/JEP.10.2012.1.2

Williams, L. A., & DeSteno, D. (2008). Pride and perseverance: The motivational role of pride. *Journal of Personality and Social Psychology*, *94*(6), 1007–1017. https://doi.org/10.1037/0022-3514.94.6.1007