



---

Year: 2019

---

## **On the psychology and economics of antisocial personality**

Fehr, Ernst ; Engelmann, Jan B ; Schmid, B ; Chumbley, J ; De Dreu, C K W

**Abstract:** How do fundamental concepts from economics, such as individuals' preferences and beliefs, relate to equally fundamental concepts from psychology, such as relatively stable personality traits? Can personality traits help us better understand economic behavior across strategic contexts? We identify an antisocial personality profile and examine the role of strategic context (the "situation"), personality traits (the "person"), and their interaction on beliefs and behaviors in trust games. Antisocial individuals exhibit a specific combination of beliefs and preferences that is difficult to reconcile with a rational choice approach that assumes that beliefs about others' behaviors are formed rationally and therefore, independently from preferences. Variations in antisocial personality are associated with effect sizes that are as large as strong variations in strategic context. Antisocial individuals have lower trust in others unless they know that they can punish them. They are also substantially less trustworthy, believe that others are like themselves, and respond to the possibility of being sanctioned more strongly, suggesting that they anticipate severe punishment if they betray their partner's trust. Antisocial individuals are not simply acting in their economic self-interest, because they harshly punish those who do not reciprocate their trust, although that reduces their economic payoff, and they do so nonimpulsively and in a very calculated manner. Antisocial individuals honor others' trust significantly less (if they cannot be punished) but also, harshly punish those who betray their trust. Overall, it seems that antisocial individuals have beliefs and behaviors based on a view of the world that assumes that most others are as antisocial as they themselves are.

DOI: <https://doi.org/10.1073/pnas.1820133116>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-182056>

Journal Article

Accepted Version

Originally published at:

Fehr, Ernst; Engelmann, Jan B; Schmid, B; Chumbley, J; De Dreu, C K W (2019). On the psychology and economics of antisocial personality. *Proceedings of the National Academy of Sciences of the United States of America*, 116(26):12781-12786.

DOI: <https://doi.org/10.1073/pnas.1820133116>

**The Dark Side of Personality: How Strategic Context and Antisocial Personality  
Interact**

Engelmann, J. B.<sup>1,2,3\*</sup>, Schmid, B.<sup>4</sup>, De Dreu, C.K.W.<sup>1,5</sup>, Chumbley, J.<sup>6</sup> & Fehr, E.<sup>7</sup>

- 1) Center for Research in Experimental Economics and Political Decision Making (CREED), University of Amsterdam, Amsterdam, The Netherlands
- 2) Amsterdam Brain and Cognition (ABC), University of Amsterdam, Amsterdam, The Netherlands
- 3) The Tinbergen Institute, Amsterdam, The Netherlands
- 4) Institute for Transport Planning and Systems (IVT), ETH Zurich, Zurich, Switzerland
- 5) Institute of Psychology, Leiden University, Leiden, The Netherlands
- 6) Jacobs Center for Productive Youth Development, University of Zurich, Switzerland
- 7) Department of Economics, University of Zurich, Zurich, Switzerland

\*Please address correspondence to: [j.b.engelmann@uva.nl](mailto:j.b.engelmann@uva.nl)

## **Abstract**

How do fundamental concepts from economics, such as individuals' preferences and beliefs, relate to equally fundamental concepts from psychology such as relatively stable personality traits? Can personality traits help us better understand economic behavior across strategic contexts? We identify an anti-social personality profile and examine the role of strategic context (the "situation"), personality traits (the "person"), and their interaction on beliefs and behaviors in trust games. Antisocial individuals exhibit a specific combination of beliefs and preferences that is difficult to reconcile with a rational choice approach that assumes that beliefs about others' behaviors are formed rationally and, therefore, independently from preferences. Variations in antisocial personality are associated with effect sizes that are as large as strong variations in strategic context. Antisocial individuals have lower trust in others unless they know that they can punish them. They are also substantially less trustworthy, believe that others are like themselves, and respond to the possibility of being sanctioned more strongly, suggesting that they anticipate severe punishment if they betray their partner's trust. Antisocial individuals are not simply acting in their economic self-interest because they harshly punish those who do not reciprocate their trust even though that reduces their economic payoff, and they do so non-impulsively and in a very calculated manner. Anti-social individuals honor others' trust significantly less (if they cannot be punished), but also harshly punish those who betray their trust. Overall, it appears that antisocial individuals have beliefs and behaviors based on a view of the world that assumes that most others are as antisocial as they themselves are.

## **Significance Statement**

Using an interdisciplinary experimental approach grounded in behavioral economics and personality psychology, we identify an anti-social personality profile and examine its role across strategic contexts. Antisocial individuals exhibit a specific combination of behaviors and beliefs: they have a high propensity to betray others' trust and believe that others are like them but, if given a punishment opportunity, they impose very harsh sanctions on those who betray their trust. More generally, antisocial individuals show beliefs and behaviors that are consistent with the assumption that most others are as antisocial as they themselves are.

/body

## **Introduction**

Personality psychology assumes that inter-individual differences are systematic and can be explained by personality traits (1). Personality traits are defined as relatively enduring patterns of thoughts, beliefs, emotions, and behaviors that reflect an individual's propensity to "respond in certain ways under specific circumstances" (2). Consistent with this definition, twin studies and recent large-scale genome-wide association studies revealed considerable genetic heritability of personality traits (3-5), and neuroimaging studies showed consistent correlations between personality traits and measures of brain structure (6-8) and brain function (9-11). Furthermore, personality traits remain relatively stable for extended periods of time (12, 13) and predict future academic achievement (14, 15), as well as personal health and wealth (16-18); for a review see (19).

Traditionally, behavior in strategic games has been examined on the basis of assumptions about the players' behavioral preferences, while the broader concept of personality traits - including thoughts and beliefs systematically tied to traits - has played little role. In fact, the typical economic analysis of games assumes that beliefs are not systematically correlated with preferences because they are formed rationally and are determined in the equilibrium - beliefs are thus not a property of the individual but a property of the group or the equilibrium the group plays.

Here we ask whether certain personality traits can help us better understand behavior in strategic games in which players face well-structured economic incentives. What is the role of variations in personality traits compared to relatively large changes in incentives for individuals' behavior? Do personality traits only play a minor role in such an environment, or do variations in personality "produce" similar behavioral effect sizes as changes in incentives? To what extent do we observe important interactions between "the person" (variations in personality traits) and "the situation" (changes in incentives)? How does the antisocial personality trait that we identify relate to economic models of social preferences (20-24)? We integrate insights from economics (19, 25) and personality theory (2, 26, 27) first to identify stable inter-individual differences in anti-social personality characteristics. Second, we examine the role of variations in anti-social personality for behavior in strategic games, and, third, we show how variations in anti-social personality interact with changes in the strategic nature of the game. We deliberately implement a large change in the strategic nature of the

game such that we can, finally, compare the effect sizes resulting from this large change in the situation with the effect size of variations in personality.

We use various versions of the trust game to study these questions. The basic trust game involves a sequential one-shot interaction between an anonymous investor and an anonymous agent who are both endowed with an identical amount of money (28). Investors decide how much of their personal monetary endowment to transfer to their agent; transfers are tripled and agents then decide how much of the available money – comprising the tripled transfer and their initial endowment – they return to their investor. In this game, both parties can earn additional money relative to their endowments, but the investor first has to trust the agent by transferring money, while the agent has to be trustworthy and transfer a sufficient amount back to the investor. The investor's transfer reflects the "behavioral trust" that the investor has in the agent and the agent's back-transfer reflects the trustworthiness of the agent, who is also free to send back nothing. The back-transfer is thus an indication of the extent to which the agent is willing to honor the investor's trust. We introduce several variations to the trust game; the most important includes the opportunity for the investor to engage in costly punishment (for details see below). We implement the possibility of sanctioning the agent very harshly because we are interested in the impact of a large "change in the situation" that can be compared to the effects of variations in antisocial personality.

### **Methods Summary and Results**

Healthy volunteers (N = 182; 98 females) filled out an online battery of personality questionnaires (see SI Appendix, section S1) and came to the laboratory approximately one week later. They participated in four different versions of the trust game and were randomly assigned either the role of the investor (Player P1; N = 90) or the agent (Player P2; N = 89). Subjects kept the same role in all four versions of the trust game. Each version of the game was played for six rounds with randomly assigned anonymous partners in each round. There were two main treatments – a binary trust game without a punishment opportunity (NPT: no punishment treatment) and a binary trust game with a punishment opportunity (PT: punishment treatment) – and two additional versions of the trust game that were included to ensure the robustness of our results (see SI Appendix, section S2 for details). To control for any potential order effects, the order in which participants faced the trust games was counterbalanced. Furthermore, we included order effect and first-round dummies in the regression analyses. To

prevent wealth effects, one round was randomly chosen to be payoff relevant at the end of the session.

In our main analyses, we focus on the differences in behavior between the PT and the NPT because this enables us to study the role of a strong situational factor (“the punishment opportunity”) and how it interacts with personality characteristics. In the NPT, the investor (“she”) and the agent (“he”) were each endowed with CHF 20, and the investor decided whether to keep the whole endowment or to transfer CHF 10 to the agent. The transfer amount was tripled and the agent then decided how much to transfer back to the investor (BT = back-transfer amount). The back-transfer was not tripled. The PT was identical to the NPT, except that after the agent made his back-transfer decision, the investor could punish the agent. For each CHF that the investor spent on punishment, she reduced the agent’s final payoff by CHF 5. Because this punishment opportunity enabled the investor to impose quite harsh sanctions on the agent, we expected substantial behavioral differences between the NPT and the PT regardless of subjects’ personality characteristics. Thus, when we later compare the relative impact of situational changes with the impact of personality characteristics, we have to keep in mind that we implemented a strong situational change.

The possible transfer (T), back-transfer (BT), and punishment (P) amounts in each game are illustrated in SI Appendix, Fig. S2. We implemented the so-called strategy method in both treatments. This means that agents in the NPT and the PT made a decision about the back-transfer for both possible cases, i.e. for an investment of CHF 0 and CHF 10. These conditional decisions by the agent were taken before he knew how much the investor had invested. Likewise, the strategy method in the PT means that the investor assigns her punishment for each of the agent’s possible back-transfers before knowing his actual back-transfer. There is substantial evidence (29) that the strategy method leads to the same qualitative behaviors as the “direct response method”. In addition, it has the advantage of providing data for those parts of the game tree that are not actually reached (e.g., we also know how much an investor would punish a back-transfer of 0 even if the agent’s actual back-transfer is high). We also measured the investor’s belief about the agent’s back-transfer in both the NPT and PT. These beliefs were elicited after the investor had made his trust decisions.

**Factor Analysis.** Given the high dimensionality of our personality data, we first employed factor analysis to reduce the data to the most essential elements and remove sources of covariance and noise before entering this data into regression analyses. An exploratory factor

analysis identified five factors (see SI Appendix, Fig. S1) including (i) Emotional reactivity (EMO), (ii) Anti-sociality (ANTI), (iii) Sensation seeking (SS), (iv) Anger (ANG), and (v) Impulsivity (IMP) (for factor loadings see SI Appendix, Table S1). We included the standardized factor scores (Bartlett scores) as explanatory variables in regression models with the following independent variables: trust (SI Appendix, Table S2), investors' beliefs about the agents' back-transfers (SI Appendix, Table S3), agents' actual back-transfers (SI Appendix, Table S4), and the investors' punishment behavior (SI Appendix, Table S5). All these regressions also include a treatment dummy coded as 1 for the PT and coded as 0 for the NPT. In addition, we controlled for a number of socio-economic characteristics, order effects, as well as for mood and stress level that were measured at the time of the experiment (SI Appendix, section S1). Since our results hold with and without these control variables, we do not discuss them in further detail. We calculated cluster-robust standard errors to account for the panel structure (within-subject design in which each individual played six rounds in each treatment). We estimated three different models for each dependent variable: Model 1 only contains the treatment dummy PT and control variables. Model 2 adds the five personality variables, and model 3 further adds the five personality x PT interactions.

**Effects of anti-sociality and punishment opportunity on behavioral trust.** For transfer levels, we find that investors armed with the option to punish agents are 15 percentage points more likely to transfer CHF 10 ( $p < 0.001$ ; see SI Appendix, Table S2; mean transfer probability PT: 74.3%; NPT: 60.7%; effect size = 0.15; see SI Appendix, Section S3 for the calculation method of standardized effect sizes). In the NPT, an increase in anti-sociality leads to a reduction in behavioral trust ( $p = 0.085$ , effect size = 0.10). For example, an individual at the 25<sup>th</sup> percentile of anti-sociality has a transfer probability that is 11.5 percentage points higher compared to one at the 75<sup>th</sup> percentile of anti-sociality (assuming average values in all other personality scores). In addition, the regression indicates a significant positive interaction between anti-sociality and the option to punish ( $p = 0.005$ , effect size = 0.11). Thus, although antisocial individuals exhibit less behavioral trust in the NPT, anti-sociality increases the impact of the punishment option on transfers substantially, raising the impact of the punishment option by 12 percentage points for a 1-SD-increase in anti-sociality. This implies (see Fig. 1A) that highly antisocial investors chose much higher trust levels in the PT compared to the NPT, and an increase in anti-sociality is even associated with a small increase in average transfers in the PT. A similar (but 50% weaker) positive interaction effect was observed for the ANG factor, but not for any other personality factor (see SI Appendix, Table S2 and Fig. S3).

The previous results suggest that antisocial individuals may have substantially less optimistic beliefs about the agents' back-transfers unless they are given the option to discipline the partner and both parties know that this option exists.

**Effects of anti-sociality and punishment opportunity on beliefs about the agents' back-transfers (“beliefs about trustworthiness”).** The role of the punishment opportunity and of antisocial personality characteristics on beliefs about back-transfers are illustrated in Fig. 1B and examined in Table S3 in the SI Appendix. The option to punish renders investors more optimistic about back-transfers ( $p < 0.001$ , effect size = 0.16), and more antisocial individuals have significantly more pessimistic beliefs about back-transfers in the absence of a punishment opportunity ( $p = 0.013$ , effect size = 0.15). A similar main effect of personality was obtained for ANG (SI Appendix, Fig. S4C). Furthermore, we find a significant and positive interaction, indicating that the impact of the punishment option on investors' beliefs increases by roughly two CHF for each 1-SD increase in anti-sociality ( $p = 0.001$ , effect size = 0.12). A negative and 40% weaker interaction effect was observed for emotional reactivity (EMO factor), but not for any other personality factor (SI Appendix, Table S3 and Fig. S4).

These results show that more anti-social investors have significantly less trust in their agents except when they know that they can punish them. This finding suggests that antisocial individuals have a particular view about human nature, i.e., they are more prone to believe that others will not honor their trust unless they can threaten to punish them.

**Effects of anti-sociality and punishment opportunity on back-transfers (“actual trustworthiness”).** If investors transfer nothing to the agent, agents' back-transfers are basically zero. The relevant question is, therefore, how the punishment opportunity and anti-sociality affect back-transfers in case the investors make a positive transfer. We find, first, that greater anti-sociality is associated with lower average back-transfer amounts ( $p = 0.001$ , effect size = 0.23; Fig. 2A and SI Appendix, Table S4). For example, an individual at the 25<sup>th</sup> percentile of anti-sociality has a back-transfer that is 44% higher compared to one at the 75<sup>th</sup> percentile of anti-sociality (CHF 13.1 versus CHF 9.1). The lower back-transfers of more antisocial agents in the NPT are not an artefact of the fact that the investors could only transfer 0 or CHF 10. In the non-binary trust game (without punishment), in which the investors could invest any amount between 0 and CHF 10, anti-sociality has an even larger effect (Fig. 2B). If investors transfer CHF 10 in this trust game, the back-transfer of subjects at the 25<sup>th</sup> percentile of anti-sociality is 64 percent higher than the transfers of those at the 75<sup>th</sup> percentile (CHF 13.7

vs. CHF 8.9). Agents who score high on the ANG factor also display lower back-transfers in the NPT, but this effect is smaller than that for anti-sociality ( $p = 0.021$ , effect size = 0.15, SI Appendix, Table S4). Conversely, the punishment option increases average back-transfers relative to the NPT by 30% for all subjects (from CHF 11.30 to CHF 14.90,  $p < 0.001$ , effect size = 0.22) and anti-sociality further increases the positive impact of the punishment opportunity on back-transfers ( $p = 0.011$ , effect size = 0.12): a 1-SD increase in anti-sociality raises the impact of the punishment opportunity on back-transfers from CHF 3.5 to roughly CHF 5.1 (i.e., by 47%). In fact, the positive interaction effect between PT and anti-sociality on back-transfers is so strong that the association between anti-sociality and back-transfers *within* the PT (see Fig. 2A) becomes rather weak. When punishment is possible, anti-social individuals are less likely to reciprocate trust, but the overall effect is small because the punishment option greatly deters them. No other personality factor shows a significant interaction with PT on back-transfers (SI Appendix, Table S4 and Fig. S5).

Why do antisocial individuals in the role of the agent respond so much more strongly to the punishment opportunity? One possibility is that anti-social individuals display different punishment patterns and – when in the role of agents – use introspection into how they would punish to form their expectations about the likelihood of being punished for low back-transfers. If this hypothesis is correct, we should observe that antisocial individuals display harsher punishments for a lack of reciprocation to positive transfers. The punishment pattern is also interesting for another reason, because those who punish only incur a material cost without reaping any reputational or material benefit from the punishment in our experimental set-up. Thus, investors who punish clearly do not maximize their own economic benefits. The punishment option therefore enables us to assess an important aspect of the nature of anti-sociality: is it merely an extreme form of selfish, payoff-maximizing behavior, or does it represent a form of malevolence in the sense that antisocial individuals display a high willingness to betray other individuals' trust while simultaneously imposing very harsh punishments on those who betray their trust, even though these punishments yield no benefits but are instead costly for them?

**Effects of anti-sociality on costly punishment.** For punishment, we estimated models that can capture a non-linear relationship between the size of the back-transfer and the punishment imposed on the agent. The reason is that the severity of the punishment may be particularly high for very low back-transfer levels. Figure 3 and Figure S7 in the SI Appendix indicate that this is indeed the case. We borrowed the candidate functions for capturing this non-linearity

from research on intertemporal choice behavior that uses non-linear least squares regression (30, 31). A model comparison revealed that a quasi-hyperbolic, double-exponential functional model best fits the data (SI Appendix, Equation S5; Fig. S7; for model comparison see Table S6). The slope coefficients for back-transfer levels (see coefficients for BT1 and BT2 in SI Appendix, Table S5) and Fig. 3 indicate that there is substantial average punishment at a back-transfer of zero that declines non-linearly until a back-transfer of CHF 25; there is little average punishment for higher back-transfers. Interestingly, highly anti-social individuals punish low back-transfers much more harshly than individuals with low levels of anti-sociality. Individuals at the 75<sup>th</sup> percentile of anti-sociality spend more than twice as much to punish back-transfers of zero compared to individuals at the 25<sup>th</sup> percentile (CHF 4.7 versus CHF 2.3, Fig. 3). Statistically, the effect of anti-sociality on punishment shows up through a positive effect on the intercept ( $p < 0.001$ , effect size = 0.35) and a significant negative interaction between anti-sociality and BT2 ( $p = 0.016$ , effect size = 0.13), which indicates the steeper decline of punishment as back-transfers increase (SI Appendix, Table S5). Finally, a similar main effect of personality was obtained for IMP (SI Appendix, Fig. S6B), but the size of the effect is approximately half as large as that for anti-sociality.

For the interpretation of the punishment behavior, it is important to keep two points in mind. First, we used the strategy method for eliciting the investors' sanctions, i.e., the investors made the punishment decision in an emotionally "cold" state in which they did not know whether their agent had in fact chosen a low back-transfer. Second, the investors decided on the punishment of an individual whom they did not know and would not meet again during the experimental session. Thus, the investors could not derive any pecuniary benefit from punishing, and they determined their punishment in a "cold", "calculative" state, assigning a punishment amount for each possible back-transfer level.

**Control analyses.** A set of control analyses are reported in the supplementary materials that check the robustness and ecological validity of the results reported above. Results from these analyses indicate that differences in emotional arousal across game contexts (see SI Appendix, Section S5, and Table S7), the limited transfer options in the binary game context (see SI Appendix, Section S6, Tables S8-S9, and Fig. S8), risk attitudes (see SI Appendix, Section S7, and Tables S10-S12), and cognitive ability (see SI Appendix, Section S8, and Table S13) likely do not drive the effects reported above.

**Situational versus personality variables.** Our set-up makes it possible to examine the relative effect sizes of changes in the situation (PT versus NPT) and changes in antisocial personality scores. We computed standardized effect sizes that make such a comparison possible in Fig. 4. The figure shows that variations in personality generally have a very similar effect size compared to the changes induced by the punishment option. In this context, it is worthwhile to mention that the PT introduces a very strong punishment opportunity that enables the investors to impose large sanctions on the agents. Therefore, if the personality variables show similar effect sizes, one may view this as a strong result. In fact, the strongest effect size emerges within the punishment treatment, where antisocial individuals impose considerably harsher sanctions on agents who provide low back-transfers (effect size = 0.35). It is also interesting that significant interactions between the situation and antisocial personality scores exist for major dependent variables – investors' transfers, investors' beliefs about back-transfers, and the agents' actual back-transfers. The identification of anti-social personality profiles thus makes it possible to understand the behavioral changes that occur when a punishment option is introduced at a much deeper level, as they interact in important ways with individuals' level of anti-sociality.

### **Discussion and Conclusions**

Personality traits are defined as relatively enduring patterns of people's thoughts, beliefs, and behaviors, i.e., thoughts and beliefs are considered a fundamental part of an individual's personality from this perspective (e.g., (32)). This view contrasts with the prevalent assumption in economics, where preferences are viewed as an individual characteristic, while beliefs are a property of the equilibrium and are often assumed to be formed rationally on the basis of the available evidence. This means that there should be no systematic relationship between what people believe about others' behavior and their own preferences. We should thus not expect a systematic correlation between individuals' social preferences and individuals' beliefs about their partner's behavior in a trust game.

Here, we used well-validated self-report measures from personality psychology to identify an antisocial personality profile that shows a systematic correlation between the beliefs of antisocial individuals and their behaviors. Moreover, no current theory of social preferences (in which pure selfishness is a special case) appears to be able to account for the belief-behavior patterns of the antisocial individuals. Antisocial individuals have high positive loadings on Machiavellianism and high negative loadings on empathy, trustworthiness, and agreeableness.

Antisocial investors in the no-punishment treatment (NPT) believe that their agents' will be considerably less likely to honor their trust compared to investors with low anti-sociality. Accordingly, highly antisocial investors also exhibit less behavioral trust than prosocial investors. Interestingly, however, if investors can punish their agents, highly antisocial investors respond to the punishment option with their beliefs and transfers much more strongly than prosocial investors: they become much more optimistic about their agents' back-transfers, and they exhibit much more behavioral trust relative to a situation without a punishment option.

Why do antisocial investors respond so differently to the existence of a punishment option? A plausible reason is that they take their own behavioral response to the punishment option into account and assume that other individuals – their agents – respond like they do. We observe, in fact, that antisocial individuals in the role of agents respond much more strongly to the punishment option by increasing their back-transfers relative to a situation without punishment. Thus, when antisocial investors put themselves into their agents' shoes and take their own likely response to the punishment option into account when predicting their agents' back-transfers, they should become more optimistic and they have reason to show more behavioral trust.

Yet, why do antisocial agents respond so strongly to the presence of a punishment option? Again, a plausible reason is that antisocial agents may introspect into how they would punish agents if they were in the role of an investor who could punish. In fact, antisocial investors indeed punish agents who provide low back-transfers much more severely compared to more prosocial investors; and they do so not in a “hot” state after their trust has actually been betrayed, but in a “cold” state in which they assign their sanctions to each possible back-transfer level ex-ante, i.e., before they know their agent's back-transfer. Therefore, if antisocial agents project their own harsh sanction – if they were an investor – to their actual investors, it makes sense for them to respond more strongly to the punishment option by raising their back-transfers. The overall picture that emerges is thus consistent with the idea that highly antisocial individuals believe that their partners in the trust game will behave similarly to how they themselves would in the other role.

No economic theory of social preferences predicts such a correlation between what antisocial individuals believe and the preferences that antisocial individuals reveal in the NPT and the PT. Note that a clean identification of behavioral preferences is only possible for agents in the NPT and for the investors at the punishment stage of the PT because the respective players

make the last move in the game in these situations and their behavior therefore does not depend on beliefs about their opponents' responses but only on their preferences.

These behavioral data clearly show that antisocial individuals are not simply selfish, because they still show significantly positive back-transfers if the investor transfers CHF 10. They spend a lot of resources on the sanctioning of those agents who make low back-transfers, although they derive no pecuniary benefit from this. Antisocial individuals are less trustworthy and less willing to honor others' trust, but they are not complete defectors. These data also imply that antisocial individuals do not simply have envious or spiteful preferences in the sense that they unconditionally value the other players' payoffs negatively, because such preferences would imply unconditional back-transfers of zero in the trust game without punishment, and they would be hard to reconcile with the fact that antisocial individuals condition their sanctions on the agents' back-transfer levels. Individuals who are simply envious or spiteful would always – regardless of the agents' back-transfer – exploit the opportunity to reduce others' material payoffs. The behavioral preferences revealed by the agents in the NPT are consistent with *weak* preferences for reciprocity (20, 23, 24) or a *weak* aversion against advantageous inequality (21). In contrast, the behavior of antisocial investors at the punishment stage is consistent with a *strong* preference for negative reciprocity (20, 23, 24) or a *strong* aversion against disadvantageous inequality (21). However, neither theories of reciprocity nor the theory of inequity aversion explains why these antisocial individuals appear to have quite different beliefs compared to less antisocial individuals. Nevertheless, their belief-behavior pattern is coherent and makes sense if these individuals assume that most others are similar to themselves, i.e., exhibit a low willingness to honor others' trust but a high willingness to punish others if their trust is not honored. In fact, this process of simulating one's own thoughts, feelings, and intentions within a hypothetical scenario to make predictions about others' behavior is referred to as *self-projection* (33), and there is considerable support from social neuroeconomics that investors utilize the neural processes that enable self-projection when making decisions in trust and similar games (34-36).

The more general lesson from our research is perhaps that beliefs may be a property of the individual as much as preferences are, and that certain belief-behavior patterns may be associated with certain personality characteristics. The psychology of personality traits may thus help in constraining the assumptions that one can or should make about what people believe about others. In addition, our research shows that variations in personality can be as important as variations in “the situation”, and that important interactions between personality characteristics and situational features exist (2, 19, 25-27).

## Materials and Methods

**Participants.** 182 volunteers from various universities in Zurich participated. Students with a background in economics and psychology were excluded to avoid potential decision biases. Three participants were removed from further analyses due to technical problems (1) and because personality scores identified them as outliers (2) (final  $N = 179$ ). Participants gave written and informed consent and all experimental procedures were approved by the ethics committee of the Institute for Empirical Research in Economics, University of Zurich.

**Procedure.** First, participants filled out an online battery of personality questionnaires several days ( $M \pm SD = 5.0 \pm 3.2$ ) before the experiment (see SI Appendix, section S1). Second, they were invited to the UZH laboratory and randomly assigned to the role of either the investor or agent. Each participant completed six rounds of each trust game treatment in pseudo-random order: the No Punishment (NPT) and Punishment Treatment (PT) outlined in SI Appendix, Fig. S1, as well as a Non-Binary Trust game (NBT) and a Direct Feedback Treatment (DT). Additional information is provided in SI Appendix, section S2.

## Acknowledgments

We gratefully acknowledge financial support from the NCCR Affective Sciences and Amsterdam Brain and Cognition (ABC). We thank Carola Hug for assistance with data collection and Joel van der Weele, Theo Offerman, and Isabel Thielmann for comments on an earlier version of this manuscript.

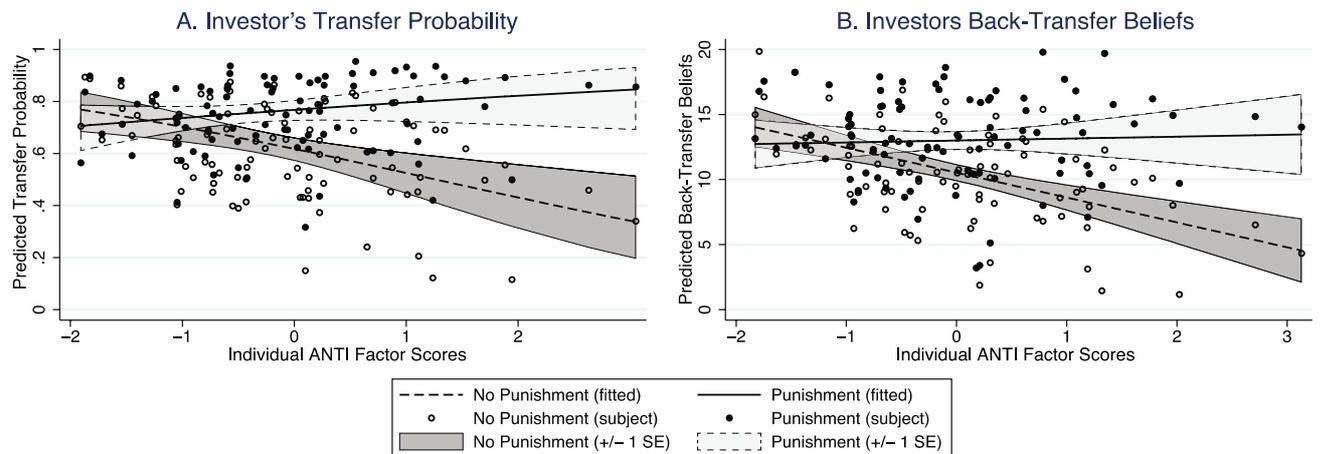
## References

1. Roberts BW, DelVecchio WF (2000) The Rank-Order Consistency of Personality Traits From Childhood to Old Age: A Quantitative Review of Longitudinal Studies. *Psychological Bulletin* 126(1):3–25.
2. Roberts BW (2009) Back to the future: Personality and Assessment and personality development. *Journal of Research in Personality* 43(2):137–145.
3. Jang KL, Livesley WJ, Vernon PA (1996) Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study. *Journal of Personality* 64(3):577–592.
4. Yamagata S, et al. (2006) Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. *Journal of Personality and Social Psychology* 90(6):987–998.
5. Lo M-T, et al. (2016) Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nature Genetics* 49(1):152–156.
6. Cohen MX, Schoene-Bake J-C, Elger CE, Weber B (2008) Connectivity-based segregation of the human striatum predicts personality characteristics. *Nat Neurosci* 12(1):32–34.
7. Cremers H, et al. (2011) Extraversion Is Linked to Volume of the Orbitofrontal Cortex and Amygdala. *PLoS ONE* 6(12):e28421.
8. DeYoung CG, et al. (2010) Testing Predictions From Personality Neuroscience: Brain Structure and the Big Five. *Psychological Science* 21(6):820–828.
9. Canli T, et al. (2001) An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience* 115(1):33–42.
10. Gray JR, et al. (2005) Affective personality differences in neural processing efficiency confirmed using fMRI. *Cogn Affect Behav Neurosci* 5(2):182–190.
11. Markett S, Montag C, Melchers M, Weber B, Reuter M (2016) Anxious personality and functional efficiency of the insular-opercular network: A graph-analytic approach to resting-state fMRI. *Cogn Affect Behav Neurosci* 16(6):1039–1049.
12. Cobb-Clark DA, Schurer S (2012) The stability of big-five personality traits. *Economics Letters* 115(1):11–15.
13. McCrae RR, Paul T Costa J (2016) The Stability of Personality: Observations and Evaluations. *Current Directions in Psychological Science* 3(6):173–175.
14. Komarraju M, Karau SJ, Schmeck RR, Avdic A (2011) The Big Five personality traits, learning styles, and academic achievement. *Personality and Individual Differences* 51(4):472–477.
15. Heckman JJ (2006) Skill Formation and the Economics of Investing in Disadvantaged Children. *Science* 312(5782):1900–1902.

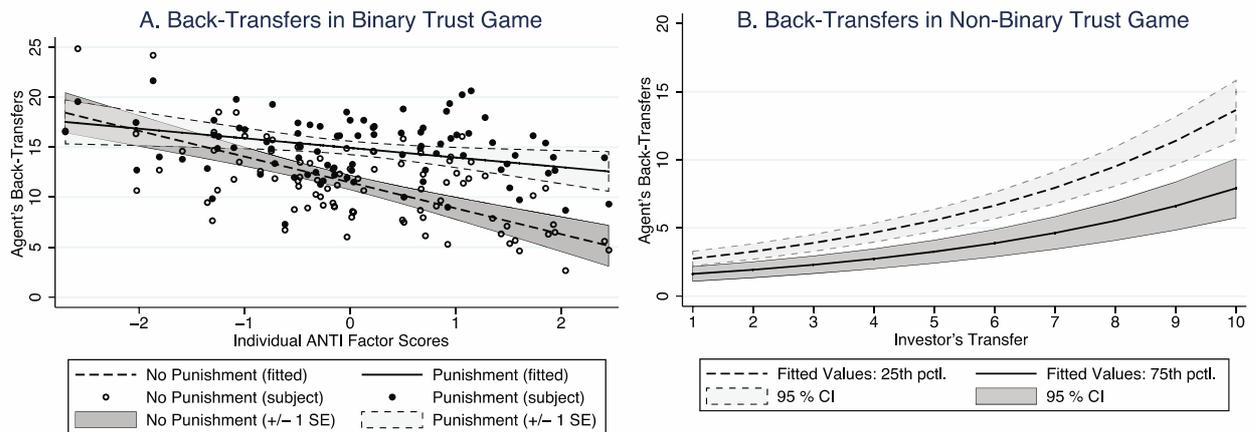
16. Moffitt TE, et al. (2011) A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences* 108(7):2693–2698.
17. Lahey BB (2009) Public health significance of neuroticism. *Am Psychol* 64(4):241–256.
18. Becker A, Deckers T, Dohmen T, Falk A, Kosse F (2012) The Relationship Between Economic Preferences and Psychological Personality Measures. *Annu Rev Econ* 4(1):453–478.
19. Almlund M, Duckworth AL, Heckman J, Kautz T (2011) Personality Psychology and Economics. *Handbook of the Economics of Education* 4:1–181.
20. Rabin M (1993) Incorporating Fairness into Game Theory and Economics. *The American Economic Review* 83(5):1281–1302.
21. Fehr E, Schmidt KM (1999) A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114(3):817–868.
22. Charness G, Rabin M (2002) Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics* 117(3):817–869.
23. Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. *Games and Economic Behavior* 47(2):268–298.
24. Falk A, Fischbacher U (2006) A theory of reciprocity. *Games and Economic Behavior* 54(2):293–315.
25. Ferguson E, Heckman JJ, Corr P (2011) Personality and economics: Overview and proposed framework. *Personality and Individual Differences* 51(3):201–209.
26. Mischel W (2004) Toward an Integrative Science of the Person. *Annu Rev Psychol* 55(1):1–22.
27. Krueger J (2009) A componential model of situation effects, person effects, and situation-by-person interaction effects on social behavior. *Journal of Research in Personality*. doi:10.1016/j.jrp.2008.12.042.
28. Berg J, Dickhaut J, McCabe K (1995) Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10(1):122–142.
29. Brandts J, Charness G (2011) The strategy versus the direct-response method: a first survey of experimental comparisons. *Exp Econ* 14(3):375–398.
30. Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. *Nat Neurosci* 10(12):1625–1633.
31. Laibson D (1997) Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics* 112(2):443–477.
32. Dweck CS (2009) Can Personality Be Changed? The Role of Beliefs in Personality and Change. *Current Directions in Psychological Science* 17(6):391–394.

33. Waytz A, Mitchell JP (2011) Two Mechanisms for Simulating Other Minds: Dissociations Between Mirroring and Self-Projection. *Current Directions in Psychological Science* 20(3):197–200.
34. Engelmann JB, Meyer F, Ruff CC, Fehr E (2019) The neural circuitry of affect-induced distortions of trust. *Sci Adv* 5(3):eaau3413.
35. Schurz M, Radua J, Aichhorn M, Richlan F, Perner J (2014) Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews* 42:9–34.
36. Mitchell JP (2009) Inferences about mental states. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1521):1309–1316

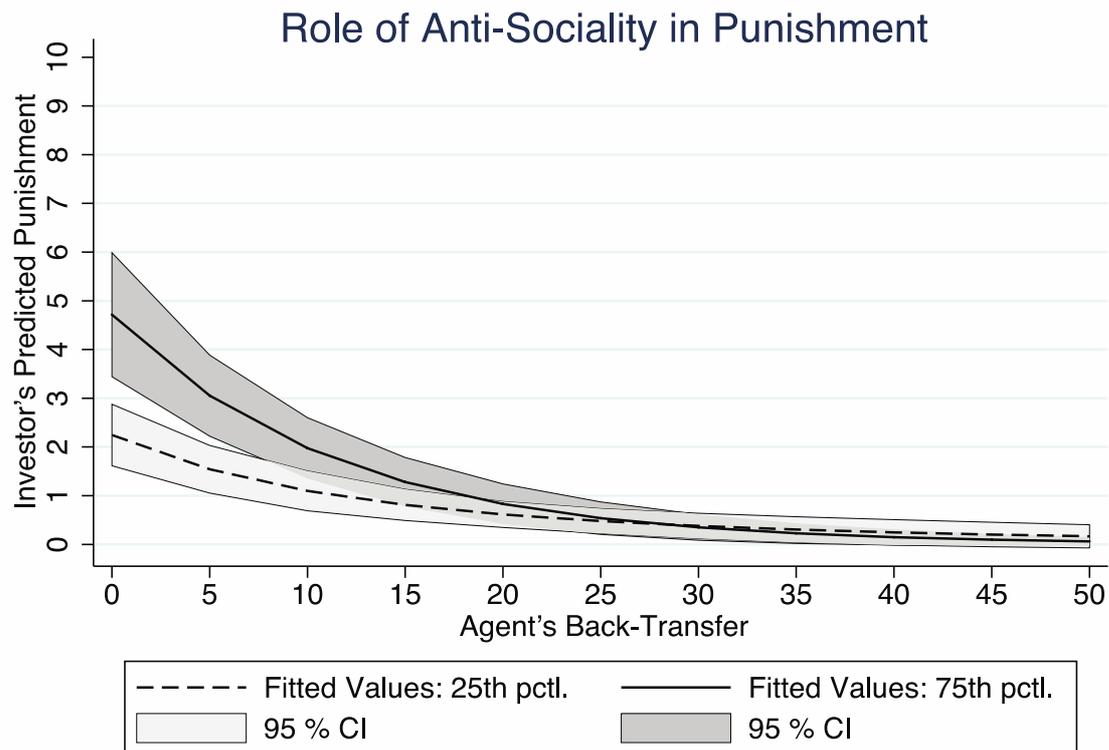
## Figures



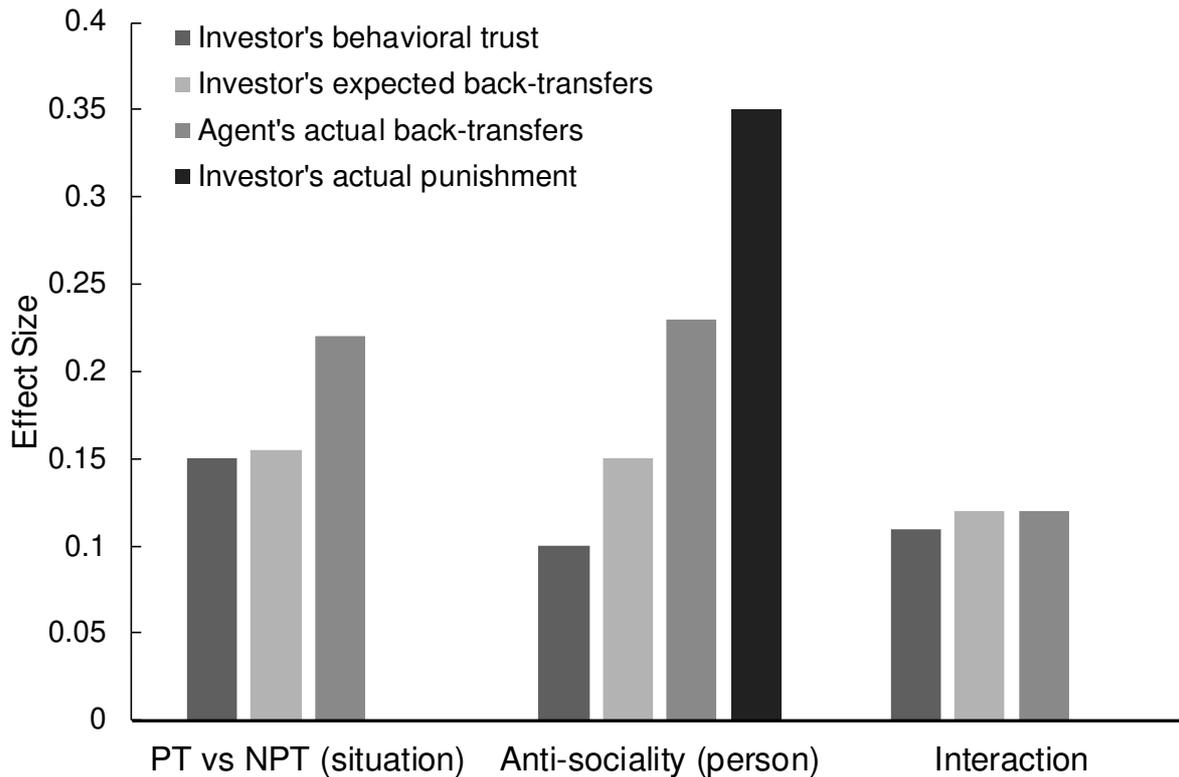
**Fig. 1: Investors' transfers (A) and beliefs about back-transfers (B) as a function of punishment condition (absent vs. present) and anti-social personality scores. Dots (N = 90) indicate individual investors' predicted probability of transferring CHF 10 (A) and their predicted beliefs about the agents' back-transfers (B) in the no punishment treatment (NPT) and the punishment treatment (PT). The predictions are based on model 3 in SI Appendix, Tables S2 (for A) and S3 (for B) and individuals' scores across all personality factors. **Solid/dashed lines:** Fitted values (A: based on model 3 in SI Appendix, Table S2; B: based on model 3 in SI Appendix, Table S3). All covariates (except for anti-sociality score and treatment (PT, NPT) were fixed at their means. *In the absence of punishment (NPT)*, a decrease in anti-sociality, e. g., from the 75<sup>th</sup> to the 25<sup>th</sup> percentile, is associated with an increase in transfer probability by 20 percent (from 56.9% to 68.4 %;  $p = 0.085$ ) and an increase in the expected back-transfer by 27 percent (from CHF 9.3 to CHF 11.8;  $p = 0.013$ ). Conversely, when the *punishment option is present (PT)*, the anti-social investors have even a somewhat *higher* transfer probability and slightly more optimistic back-transfer expectations. This is due to a relatively strong interaction between anti-sociality and the punishment option ( $p = 0.005$ , SI Appendix, Table S2): for every one-SD increase in anti-sociality, the PT effect on transfers (and expected back-transfers) increases by 12 percentage points (and CHF 2.1). None of the other personality traits has a significant interaction with PT for transfers (SI Appendix, Table S2, Fig. S3). For beliefs, only emotional reactivity shows a significant interaction with PT (SI Appendix, Table S3, Fig. S4), such that individuals with low emotional reactivity expect significantly larger back-transfers in PT.**



**Fig. 2: Agents' back-transfers as a function of punishment condition (absent vs. present) and antisocial personality scores in the binary trust game (A). In (B) agent's back-transfers in the non-binary trust game (NBT) are shown as a function of investors' transfer levels (that could vary between 0 and 10) and for agents at the 25<sup>th</sup> and the 75<sup>th</sup> percentile of anti-sociality. Dots (N = 89) indicate individual agents' predicted back-transfer in the no punishment treatment (NPT) and the punishment treatment (PT). The predictions result from model 3 in Table S4 in the SI Appendix by inserting for each agent the individual personality factor scores. **Solid/dashed lines:** Fitted values (A: based on model 3 in SI Appendix, Table S4; B: based on model 3 in SI Appendix, Table S9). All covariates (except for anti-sociality score and treatment (PT, NPT)) were fixed at their means. In the *absence of punishment (NPT)*, a decrease in the anti-sociality from the 75<sup>th</sup> to the 25<sup>th</sup> percentile is associated with an increase in average back-transfers by 44% in the binary trust game (A: from CHF 9.1 to CHF 13.1;  $p = 0.001$ ); the increase in back-transfers at a transfer of CHF 10 is even 64% in the non-binary trust game with continuous investor transfers (B: from CHF 8.9 to CHF 13.7;  $p = 0.001$ ; SI Appendix, Table S9). In contrast, the back-transfer gap between highly antisocial (75<sup>th</sup> percentile) individuals and those at the 25<sup>th</sup> percentile is only 11% in the *presence of punishment (PT)* (A: from CHF 14.00 to CHF 15.5), which is due to the much stronger response of antisocial individuals to the punishment threat: for every one-SD increase in anti-sociality, the PT effect on back-transfers increases by roughly CHF 1.6 (SI Appendix, Table S4;  $p = 0.011$ ). None of the other personality traits has a significant interaction with PT (SI Appendix, Table S4 and Fig. S5).**



**Fig. 3: Expenditures of investors (who transfer CHF 10) on punishment for different back-transfer levels for investors with a high (75<sup>th</sup> percentile) and a low (25<sup>th</sup> percentile) anti-sociality score (N = 80). Solid/dashed lines: Fitted values are based on model 3 in the SI Appendix, Table S5. All covariates were fixed at their means. Highly antisocial investors spend considerably more on the punishment of low back-transfers than less antisocial investors (SI Appendix, Table S5, model 3;  $p < 0.001$ ). Those at the 75<sup>th</sup> percentile of anti-sociality spend 110% more on punishment than those at the 25<sup>th</sup> percentile at a back-transfer of zero (CHF 4.72 compared to CHF 2.25). Punishment is generally low beyond back-transfers of CHF 25 and identical across levels of anti-sociality.**



**Fig. 4: Standardized effect sizes (see SI Appendix, Section S3 for the calculation method) of changes in the situation (Punishment Treatment (PT) versus No Punishment Treatment (NPT)), variations in the antisocial personality scores and their interaction on investors' behavioral trust, investors' beliefs in back-transfers, and agents' actual back-transfers.** Furthermore, we also show the effect size of the anti-sociality score on investors' punishment behaviors. Variations in antisocial personality scores have effect sizes that are quite comparable to the effect size of a punishment option that enables investors to impose severe sanctions on agents (for every CHF the investor spends on punishment, the agent loses CHF 5). For behavioral trust, the effect size of the situational variable (PT vs. NPT) is somewhat larger than the effect size for anti-social personality, while the effect size of the situational variable is almost identical to the personality variable for investors' expected back-transfers and agents' actual back-transfers. However, by far the biggest effect emerges for the personality effect on investors' punishment. In addition, we observe significant interactions between the situational and the personality variable.