Theory of Mind and the Agreeableness-Antagonism Dimension:

Differential Associations with Callousness, Aggression, and Manipulativeness

Scott D. Blain[1]*, Aisha L. Udochi[2], Timothy A. Allen[3], Muchen Xi[2], & Colin G. DeYoung[2]

University of Michigan Department of Psychiatry[1], University of Minnesota – Twin Cities Department of Psychology[2], University of Pittsburgh Department of Psychiatry[3]

*Corresponding Author:

Scott D. Blain

University of Michigan Department of Psychiatry

Rachel Upjohn Building 2742

4250 Plymouth Road

Ann Arbor, MI 48109

(270) 287-8688

blscott@med.umich.edu

Author Note:

SDB and ALU were supported by National Science Foundation Graduate Research Fellowships, during the completion of this work. Additional support was provided by Grant No. K01-MH123915 to TAA.

Preliminary results were presented by SDB at various conferences and campus talks, including a biopsychology colloquium at Ruhr University Bochum (2019), the Association for Research in Personality (2019), International Society for the Study of Individual Differences (2019), and Society for Personality and Social Psychology (2021). Finally, this work was used in part to fulfill Ph.D. degree requirements for SDB and was included in the dissertation "Individual Differences in Social Cognition and Behavior: A Personality Psychology Framework."

Analytical scripts, study data, a list of all procedures and measures included in this study, and a list of references also utilizing the same broader dataset (currently none but other manuscripts are in preparation) will be made publicly available at the Open Science Framework (https://osf.io/3qwjr/?view_only=8ff037d28a5247c585c780faedb534bb). This article was also published as a digital preprint, in efforts to facilitate open science.

Abstract

Theory of Mind (ToM) refers to how we identify and understand the mental states of others. ToM abilities vary with dimensions of normal-range personality and can be seriously impaired in a number of mental disorders, particularly those related to the Antagonism domain. The current study used a multi-task design to examine how ToM relates to Agreeableness-Antagonism subfactors, replicating and extending previous work. Participants (N = 335) completed self-report measures of the Big Five, empathy, and personality pathology, as well as tasks spanning mental state attribution, affect recognition, and mentalizing. Exploratory structural equation modeling was used to assess the impact of Agreeableness-Antagonism subfactors on ToM. A three-factor structure was derived for Agreeableness-Antagonism, with factors corresponding to Compassion-Callousness, Pacifism-Aggression, and Honesty-Manipulativeness. While higher Aggression and lower Compassion predicted worse ToM ability, higher Manipulativeness predicted better ToM ability. Findings replicate and extend work suggesting differential relations of specific Agreeableness-Antagonism subfactors with social cognition. We discuss our results with a focus on the importance of dimensional psychopathology models and facet-level research.

Key Words: Social Cognition, Personality, Theory of Mind, Dark Triad, Emotional Intelligence

Theory of Mind and the Agreeableness-Antagonism Dimension: Differential Associations with Callousness, Aggression, and Manipulativeness

When navigating social interactions, humans rely upon a variety of social cognitive processes—including the ability to perceive emotions and empathize with others (Barrett et al, 2011; Singer & Klimecki, 2014). One important social cognitive process is known as *theory of mind (ToM)* or *mentalizing*, which refers to a person's ability to recognize, understand, and utilize the thoughts, feelings, and beliefs of other people (Premack & Woodruff, 1978). People vary in their ToM capacity, and research in psychopathology has consistently reported that social cognitive abilities covary with a broad range of diagnoses and dimensions—including Autism Spectrum Quotient, Psychoticism, and Antagonism (Baron-Cohen et al., 1986; Bora & Pantelis, 2013; Dolan & Fullam, 2004; Krueger & Markon, 2014; Preißler et al., 2010). Better ToM is also associated with functional outcomes such as increased social competence (Liddle & Nettle, 2006; Jensen-Campbell et al., 2002) and reduced antisocial behavior (Mohr et al., 2007; Meier et al., 2006). Given that the field of personality psychology seeks to provide comprehensive taxonomies and explanatory models for understanding individual differences in human cognition and behavior, looking to existing models of normal-range personality variation provides one promising avenue for better understanding ToM and its potential correlates.

Social Cognition and Personality

The *Big Five* personality traits capture five broad dimensions of personality that can comprehensively organize most personality traits and descriptors (John et al., 2008). Each of the Big Five has been linked to individual differences in specific psychological processes and associated self-regulatory functions (DeYoung, 2015; DeYoung & Blain, 2020). For instance, Extraversion is related to reward responsivity and Openness/Intellect is related to information

processing and pattern sensitivity (Blain, Grazioplene et al., 2020; Blain, Longenecker et al., 2020; Blain, Sassenberg et al., 2020; DeYoung et al., 2012; 2014; Lucas et al., 2000; Smillie et al., 2012; 2019). ToM and social cognition appear most related to the Big Five domain of Agreeableness. Agreeableness has been associated with many of the same functional outcomes as ToM, including social competence and social network size (Liddle & Nettle, 2006; Jensen-Campbell et al., 2002; Meier et al., 2006). Research directly examining the relation between Agreeableness and ToM has also begun to emerge, and there is some evidence for a positive correlation between the two constructs, such as findings presented by Nettle & Liddle (2008). Further insight into the individual difference correlates of ToM has been gained by simultaneously examining its association with normal-range and pathological personality traits.

The Big Five dimensions are very similar to the dimensions that emerge from patterns of covariation in symptoms of psychopathology—including not only personality disorders but other disorders too (DeYoung & Krueger, 2018; Kotov et al., 2010; Kotov et al., 2017). Many psychiatric symptoms can be described as risky or maladaptive variants of behaviors described by normal personality variation (DeYoung & Krueger, 2018). For instance, maladaptively low Agreeableness has been labeled "Antagonism" (Gore & Widiger, 2013; Suzuki et al., 2015). Individual differences in ToM ability have been linked to both Agreeableness and Antagonism, including factors representing the shared variance of Agreeableness and Antagonism scales (Allen et al., 2017; Nettle & Liddle, 2008). Importantly, Agreeableness and Antagonism, like all domain-level traits, can be broken down into a variety of lower-order traits, which are typically labeled aspect and facets (DeYoung et al., 2007). For instance, Agreeableness can be decomposed into the two aspects of *Compassion* and *Politeness*, which describe peoples' ability to empathize with others and peoples' ability to control socially unacceptable behaviors (e.g.,

expressions of aggression), respectively (DeYoung et al., 2007). Facets of Antagonism include callousness, aggression, and manipulativeness (Gore & Widiger, 2013; Krueger et al., 2013).

Allen et al. (2017) directly examined differential associations between ToM ability and lower-order subfactors of the Agreeableness-Antagonism domain. Allen et al. (2017) found that ToM was positively correlated with Compassion and negatively correlated with Politeness. Subsequent analyses of multiple Agreeableness and Antagonism facet-level scales suggested that Politeness could be subdivided into two subfactors that differentially predicted ToM ability. Like the Compassion aspect, a Non-aggression or Pacifism subfactor positively predicted ToM, but an Honesty subfactor negatively predicted ToM. Nonetheless, these findings regarding subfactors were discovered in *post hoc* analyses and warrant replication using a confirmatory framework and additional measures of ToM ability. In the current research, we attempted to replicate and extend the work of Allen et al. (2017), further examining whether ToM ability was related to these three Agreeableness-Antagonism subfactors of Compassion-Callousness, Honesty-Manipulativeness, and Pacifism-Aggression.

Reliability, Multi-task Designs, and Latent Variable Modeling

To justify claims regarding the underlying associations among constructs—for example, Agreeableness and social cognition—we must first be able to assess each of those constructs, individually, in a way that is reliable and valid. Concerns of reliability and validity are especially important when using behavioral tasks, as even tasks that can detect robust effects at the group level (e.g., tests of implicit bias or self-regulation) often fail to produce reliable measurement of individual differences (Hedge et al., 2018; Enkavi et al., 2019a; Schnabel et al., 2008). Fortunately, questionnaire measures of personality and tests of general or social cognitive ability tend to have better reliabilities than many of the measures commonly used in other areas of

psychology (Hedge et al., 2018; Morrison et al., 2019; Pinkham et al., 2018; Vellante et al., 2013). Regardless of their reliability, however, single-task performance-based indicators are often limited in their scope and measure constructs more narrowly than those they purport to represent (Apperly, 2012; Blain, Longenecker et al., 2020). Performance on any given task is influenced by a number of task-specific factors but using multi-indicator designs can allow us to move toward measuring constructs more reliably as what is shared across multiple tasks, thereby avoiding underestimation of true effect sizes (Blain, Longenecker et al., 2020; Campbell & Fiske, 1959; Eisenberg et al., 2019; Enkavi et al., 2019a; 2019b; Nosek & Smythe, 2007).

We can further increase our ability to reliably measure constructs such as ToM ability and estimate their associations with other variables by using latent variable methods, such as structural equation modeling (SEM), which models the prediction of latent variables by other latent variables. Latent variables represent the shared variance of multiple measured (or *manifest*) variables (Schumacker & Lomax, 2004). For example, a latent social cognitive ability variable might be modeled as the shared variance of accuracy scores across different social cognition tasks. Assessing variables of interest at the latent level allows for more robust conclusions, as latent variables capture only the shared variance of their indicators, thereby eliminating error variance and more accurately capturing variability in the underlying constructs of interest (Keith, 2006). Modeling social cognitive ability as the shared variance in performance across tasks should give a better representation of true variance in social cognitive ability by factoring out unique task variance (which includes a combination of task-specific variance and error).

The Current Study

We hoped to replicate and extend the findings from Nettle & Liddle (2008) and Allen et

al. (2017) by analyzing the relation of ToM ability and individuals' trait levels along the Agreeableness-Antagonism continuum. We further break down the personality hierarchy to explore how Agreeableness-Antagonism subfactors (i.e., Compassion-Callousness, Pacifism-Aggression, and Honesty-Manipulativeness) relate to ToM.

First, we hypothesized that accuracy scores from multiple tests of ToM ability would be positively correlated with one another and would map onto a single latent factor, producing a well-fitting measurement model. Mirroring the pattern of findings from Allen et al. (2017), we also hypothesized that when computing an oblique, three-factor solution while factor analyzing multiple self-report measures of Agreeableness and Antagonism, dimensions would emerge that corresponded to Compassion-Callousness, Honesty-Manipulativeness, and Pacifism-Aggression. We hypothesized that ToM accuracy (modeled as a latent variable and as scores from individual tasks) would positively correlate with subfactors for Compassion and Pacifism and negatively correlate with Honesty. In an effort to more clearly replicate previous research, we also included models predicting performance on the individual tasks, in addition to our model using a latent variable for ToM accuracy.

Although the work of Allen et al. (2017) showed support for three subfactors of Agreeableness-Antagonism that differentially predict ToM abilities, their analyses were *post hoc* and warrant replication. Further, our study utilizes a broad battery of Agreeableness-Antagonism facet scales, a multitask design, and an exploratory structural equation modeling (ESEM) analytical approach—all of which are advantages over previous work done on this topic.

Method

Participants

Participants were recruited via a combination of Qualtrics panels and from the campus of

a large public research university in the Midwestern US as part of a study on social cognition, personality, and psychopathology. No explicit exclusion criteria for psychopathology were implemented in an attempt to sample a broad, representative range of pathological and normal personality variation from the general population. The original sample consisted of 389 individuals, but 54 individuals were excluded for having high amounts of random or invalid responding, leaving a total valid sample of 335. Participants ranged from 18 to 75 in age (M = 26.4, SD = 13.6). There were 267 females (79.7%), 67 males (20%), and 1 intersex individual (0.3%). In terms of race/ethnicity, 235 participants identified as White or Caucasian (70.1%), 59 as Asian or Pacific Islander (17.6%), 7 as Black or African American (2.1%), 4 as Latino or Hispanic (1.2%), and 30 as multiracial or other (9.0%). 283 participants were native English speakers (84.5%).

Participants reviewed an online informed consent document before beginning the study, then completed an online battery of questionnaires and behavioral tasks, including self-report measures of demographics, personality, psychopathology, and social functioning and several tests of social cognition. All protocols were approved by the Institutional Review Board.

Self-Report Measures

Big Five Aspect Scales

The Big Five Aspect Scales (BFAS; DeYoung et al., 2007) is a 100-item questionnaire that assesses the Big Five personality domains, including two component aspects for each of the Big Five. Each aspect is measured by a total of ten items, including a combination of standard and reverse-coded items. Participants answered each question using a five-point Likert scale ranging from 1 (“Strongly disagree”) to 5 (“Strongly agree”). The current study used the two Agreeableness aspects scales, measuring Compassion and Politeness.

Computer Adaptive Test of Personality Disorders: Static Form

The Computer Adaptive Test of Personality Disorders: Static Form (CAT-PD SF; Simms et al., 2011; Wright & Simms, 2014), a selection of 212 items from the 1366-item CAT-PD, is a measure that assesses 33 maladaptive personality traits (e.g., Callousness and Manipulativeness) that can be grouped into five broad categories similar to the Big Five (i.e., Negative Emotionality, Detachment, Antagonism, Disconstraint, and Psychoticism). Participants rated items on a 5-point scale ranging from 1 ("very untrue of me") to 5 ("very true of me"). The current study used the Antagonism-related scales of Callousness, Manipulativeness, Hostile Aggression, and Domineering.

Externalizing Spectrum Inventory Brief Form

The Externalizing Spectrum Inventory Brief Form (ESI-BF; Patrick et al., 2013) is a shortened version of the 415-item ESI. This 160-item questionnaire assesses general Disinhibition, Substance Abuse, Callous Aggression, and 23 lower-order facets of the externalizing spectrum. Participants rated each item on a 4-point scale, with higher scores corresponding to greater agreement with the item. The current study used the lower-order facet scales most strongly correlated with Antagonism and Agreeableness, including Physical Aggression, Relational Aggression, Destructive Aggression, Fraud, Theft, Empathy, and Honesty.

Interpersonal Reactivity Index

The Interpersonal Reactivity Index (IRI; Davis et al., 1980) is a 28-item questionnaire that assesses individual differences in empathy across four dimensions—Empathic Concern, Fantasy, Personal Distress, and Perspective Taking. Each dimension is measured by a 7-item subscale, and participants rated items on a 5-point Likert scale ranging from 0 ("does not

describe me well") to 4 ("describes me very well"), with a combination of standard and reverse-scored items. In the current study, we only used the Empathic Concern scale (i.e., other-oriented emotions centered on helping people in need) because it is the IRI dimension most strongly associated with Agreeableness (Melchers et al., 2016).

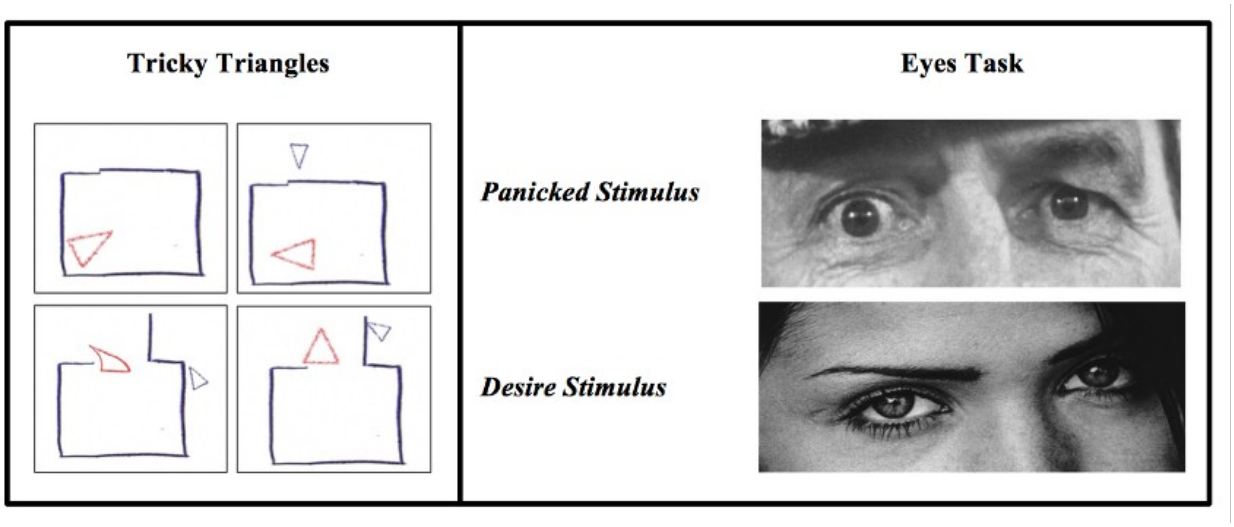
Personality Inventory for DSM-5

The Personality Inventory for DSM-5 (PID-5; Krueger et al., 2012) is a 220-item questionnaire that assesses 25 maladaptive personality trait facets that, like the CAT-PD, can be grouped into five categories (i.e., Antagonism, Detachment, Disinhibition, Negative Affect, and Psychoticism). Facet scales range from 4 to 14 items and are rated on a 4-point scale ranging from 1 ("very false or often false") to 4 ("very true or often true"). The present study used three of the facets belonging to the Antagonism domain, including Callousness, Deceitfulness, and Manipulativeness.

Theory of Mind Tasks

Participants also completed three tests of social cognition, including a triangle animation task in which participants labeled animations as random, physical, or social (Abell et al., 2000), a mentalizing stories task in which participants had to answer questions about characters' factual and social knowledge (Stiller & Dunbar, 2007), and a perceptual ToM task using eye stimuli (Baron-Cohen et al., 1997). Example stimuli from the triangles and eyes tasks are pictured in Figure 1.

Figure 1.



Note. Social cognition tasks

Theory of Mind Vignettes

The ToM vignette task (Stiller & Dunbar, 2007) comprises five short stories depicting social situations. Each story describes a social interaction involving multiple characters. Participants read each story, after which they answered five ToM questions and five memory questions pertaining to the story. All questions are in true-false format. Memory questions are designed to measure the participants' ability to retain the factual contents of the story, and the number of facts that the participant must retain varies from two to six in each question. Performance on memory questions within the task can be used as a covariate to ensure that any associations with variables of interest are due to participants' ToM ability rather than their memory for the details of the story. ToM questions required that the participant reason, or infer, a character's perspective in the story. Questions vary across five levels of difficulty, with each successive level requiring the participant to track an additional character or level of perspective. For example, in second-level questions, participants tracked their own mental state and the mental state of one character (e.g., "John wanted to go home after work"). In fourth-level questions, participants tracked the mental state of three characters (e.g., "John thought that Penny

knew what Sheila wanted to do"). To assess performance on the task, we adopted the procedure used by Nettle and Liddle (2008) and Allen et al. (2017) and computed simple sums of correct responses to memory questions and ToM questions for each participant.

Tricky Triangles Task

In the triangles task (Abell et al., 2000), participants are presented with a series of computerized animations of shapes interacting in a way that was random, physical, or social. In the random condition, the shapes did not interact with each other, but rather moved around purposelessly (e.g., bouncing or drifting). In the physical condition, the shapes moved in a goal-directed manner without invoking ToM or mentalizing (e.g., fishing or swimming). In the social condition, shapes enacted a social sequence, such as coaxing, seducing, or mocking. Participants were tasked with indicating whether each animation was random, physical, or social in nature, then scored for their accuracy in correctly categorizing each animation in a series of 22 clips.

Reading the Mind in the Eyes Task

The eyes task (Baron-Cohen et al., 2001) consists of 36 grey-scale photos of people taken from magazines. These photos were cropped and rescaled so that only the area around the eyes could be seen. Each photo was accompanied by four mental state terms, from which the participant was instructed to choose the word that best described what the person in the photo was thinking or feeling. Only one of the four items was correct (as judged by consensus from an independent panel of judges in the initial psychometric study). Participants were scored for their accuracy across all 36 stimuli.

Analyses

Descriptive statistics were calculated for all task performance and personality variables. Cronbach's α (Cronbach, 1951) and ω_t (McDonald, 1999; Revelle & Condon, 2019) were

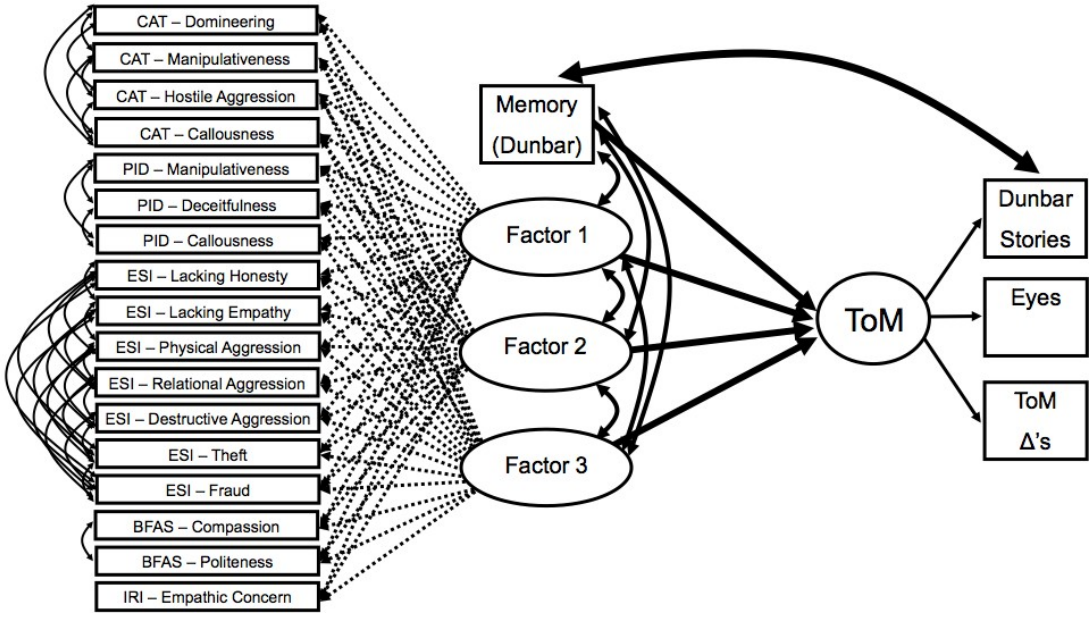
computed to assess internal consistency reliability. *Mplus* was then used for latent variable modeling (Muthén & Muthén, 2017); all models were estimated using full-information, robust maximum likelihood estimation (MLR). First, we computed a single-factor confirmatory factor analytic (CFA) model using accuracy scores from the three ToM tasks to examine how well these tasks represent a single coherent latent variable. To further assess whether a single factor model was appropriate for explaining shared variance across tasks, ω_t was also computed.

Next, Exploratory Structural Equation Modeling (ESEM) was implemented. Exploratory Agreeableness-Antagonism subfactors were derived from relevant BFAS, IRI, ESI-BF, PID-5, and CAT-PD SF subscales. First, we conducted a Velicer's minimum average partial (MAP) test (O'Connor, 2000) to see how many factors were empirically suggested in this data. Then, a three-factor solution was computed, given our hypotheses and aims to conceptually replicate and extend the work of Allen et al. (2017). We used factoring with a constrained oblimin rotation (γ = -.80). Relative to traditional structural equation modeling (SEM) with confirmatory factors, computing exploratory factors better accounts for the nontrivial cross-loadings of indicators (Asparouhov & Muthén, 2009). ESEM also allows for more accurate model estimation vs. simple use of observed factor score estimates in SEM. Nonetheless, results for models in the current study were substantively equal if factor score estimates derived using exploratory factor analysis followed by the regression method were used (rather than full ESEM).

Residual covariances of subscales from the same questionnaire (e.g., BFAS Politeness and Compassion) were freely estimated, which resulted in better fit vs. constraining these covariances to zero (ΔS-B $\chi^2 = 218.3$, $p < .001$). Additionally, the residual covariance between memory accuracy and ToM accuracy from the vignette task was freely estimated, resulting in significantly improved fit (ΔS-B $\chi^2 = 23.4$, $p < .001$).

Subsequently, we predicted latent ToM accuracy (shared variance of accuracy across the three tasks) from the Agreeableness-Antagonism factors, allowing predictors to correlate and including performance on the memory questions from the vignette task as an additional, correlated predictor variable. The full ESEM model is presented in Figure 2.

Figure 2.



Note. Exploratory structural equation model of Agreeableness-Antagonism factors and ToM

Satorra-Bentler adjusted fit indices were computed and 95% confidence intervals (with standard errors derived using the Huber-White sandwich estimator) were estimated for the path coefficients from predictor variables to latent ToM accuracy (Huber, 1967; Muthén & Muthén, 2017; Satorra, & Bentler, 2001; White, 1980). Finally, for visualization purposes, factor score estimates were computed for Agreeableness-Antagonism and ToM latent variables, using the regression method. ToM factor scores were residualized for scores on the memory condition of the vignette task and standardized, then plotted against standardized factor scores for each of the Agreeableness-Antagonism factors (residualizing for variance in the other two Agreeableness-Antagonism factors).

In a final model, which sought to more directly replicate previous results, we examined the effects of Agreeableness-Antagonism factors on observed accuracy scores for each of the individual ToM tasks (controlling for performance on the memory conditioning from the story task). These models are presented in our supplement. Results from models that did not freely estimate residual covariances of variables from the same task/questionnaire are reported in our supplemental methods and results.

Results

Descriptive statistics are presented in Table A1. Performance was generally high for the behavioral tasks, but variables showed no prominent ceiling effects. Several of the personality and task performance variables showed moderate skewness; thus, analytical methods robust to non-normality were used (i.e., MLR estimation implemented with MPLUS). Values of ω_t and α indicated acceptable internal consistency for the majority of questionnaire and task variables.

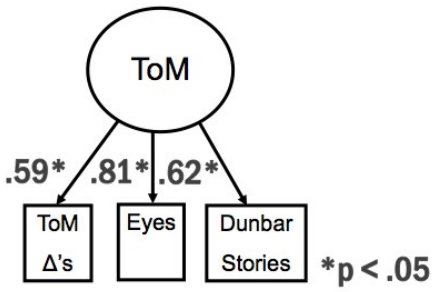
Table 1. Descriptive statistics

	Mean (SD)	Skewness	[Minimum, Maximum]	α	ω_t
Dunbar ToM	19.4 (3.2)	-1.1	[8, 25]	.66	.77
Dunbar Memory	22.5 (3.7)	-1.6	[10, 26]	.81	.85
Eyes Task	24.3 (5.7)	-0.8	[5, 34]	.77	.80
Triangles	12.6 (3.3)	-0.5	[3, 19]	.62	.68
BFAS Compassion	4.0 (0.6)	-0.4	[2.4, 5.0]	.86	.86
BFAS Politeness	3.8 (0.6)	-0.6	[2.1, 5.0]	.77	.77
IRI Empathic Concern	3.8 (0.7)	-0.2	[1.9, 5.0]	.77	.78
PID Callousness	1.4 (0.5)	2.0	[1.0, 3.8]	.91	.94
PID Deceitfulness	1.6 (0.6)	1.0	[1.0, 3.7]	.88	.89
PID Manipulativeness	1.8 (0.8)	0.6	[1.0, 4.0]	.84	.85
CAT Callousness	1.8 (0.8)	1.1	[1.0, 4.4]	.91	.91

CAT Domineering	2.1 (0.8)	0.7	[1.0, 5.0]	.85	.85
CAT Hostile Aggression	1.6 (0.8)	1.7	[1.0, 5.0]	.93	.93
CAT Manipulativeness	1.7 (0.7)	1.2	[1.0, 4.3]	.87	.87
ESI Theft	1.3 (0.7)	2.1	[1.0, 4.0]	.90	.91
ESI Fraud	1.3 (0.7)	2.2	[1.0, 4.0]	.89	.90
ESI Honesty	1.9 (0.6)	0.4	[1.0, 4.0]	.76	.77
ESI Physical Aggression	1.4 (0.6)	2.1	[1.0, 4.0]	.91	.92
ESI Destructive Aggression	1.3 (0.5)	2.6	[1.0, 4.0]	.93	.93
ESI Relational Aggression	1.4 (0.6)	1.6	[1.0, 4.0]	.89	.89
ESI Empathy	1.6 (0.5)	0.9	[1.0, 3.2]	.88	.89

Bivariate correlations are presented in Table A1. Across measures, Agreeableness variables positively predicted task performance while Antagonism variables negatively predicted task performance. The CFA model showed that accuracy scores on the three ToM tasks loaded onto a single latent variable (Figure 3). Factor loadings were moderately high, with performance on the eyes task being the strongest. Fit statistics for all models are presented in Table 2. The CFA model was just identified, so meaningful model fit evaluation based on standard fit indices was not possible. Nonetheless, there was evidence that a substantial portion of variance could be explained by a single underlying social cognitive ability factor (ω_t = .71).

Figure 3.



Note. Model of ToM tasks

Table 2. Fit statistics for structural equation models

Models	RMSEA	95% C.I.	SRMR	S-B $\chi 2$	*p*	CFI	TLI
1. ToM	.000	[.000, .000]	.000	0.00	< .001	1.0	1.0
2. ToM, Mem, and Antagonism Factors	.060	[.050, .069]	.023	262.4	< .001	.98	.96
3. Individual Tasks, Mem, and Antagonism Factors	.061	[.051, .071]	.023	255.1	< .001	.98	.96
S1. ToM, Mem, and Antagonism Factors (constrained residual covariances)	.088	[.081, .097]	.042	550.7	< .001	.94	.91
S2. Individual Tasks, Mem, and Antagonism Factors (constrained residual covariances)	.087	[.078, .095]	.028	506.0	< .001	.94	.91

Next, we used ESEM to identify latent factors from Agreeableness and Antagonism scales. Consistent with the notion that Agreeableness can be separated into two correlated aspects and with previous work using a similar set of Agreeableness-Antagonism scales (Allen et al., 2017), conducting a Velicer's MAP test indicated the presence of two factors across the 17 scales. Correlations between these two extracted factors and relevant BFAS variables showed that the first factor approximated low Politeness (r = -.66, p < .001) and the second factor strongly resembled Compassion (r = .79, p < .001).

Since we were interested in parsing subfactors within Politeness and replicating the discriminant validity of these two factors in predicting ToM (Allen et al., 2017), we then chose to extract three factors. Factor loadings for the three-factor solution are presented in Table 3.

Table 3. Factor loadings of Agreeableness-Antagonism scales on three exploratory factors

Scale	Aggression	Manipulativeness	Compassion
CAT – Domineering	.21	**.57***	-.18*
CAT – Manipulativeness	**.47***	.46*	-.14*

PID – Manipulativeness	.24	**.60***	-.02
PID – Deceitfulness	.28	**.61***	-.12*
ESI – Honesty	-.02	**-.39***	.16
BFAS – Politeness	-.07	**-.51***	.26*
ESI – Relational Aggression	**.53***	.39*	-.10*
ESI – Physical Aggression	**.82***	.03	-.03
ESI – Destructive Aggression	**.90***	-.03	-.03
ESI – Theft	**.78***	.13	.05
ESI – Fraud	**.70***	.20	-.03
CAT – Hostile Aggression	**.82***	.05	.14*
PID – Callousness	**.66***	.06	.31*
CAT – Callousness	.40*	.16*	**-.52***
ESI – Empathy	-.29*	-.04	**.70***
BFAS – Compassion	-.02	-.08	**.78***
IRI – Empathic Concern	-.05	-.10	**.71***

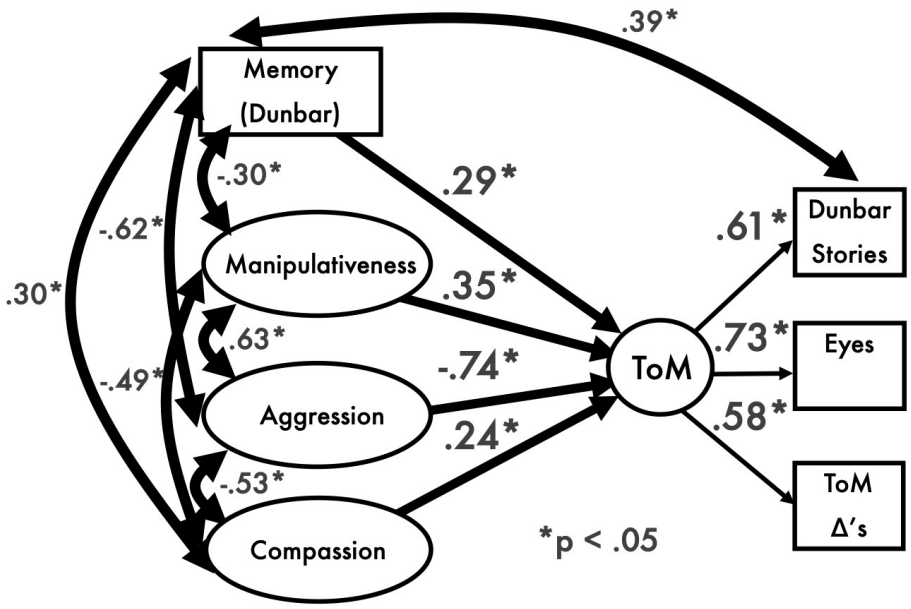
Note. *$p < .05$ (based on the z-distribution and standard errors computed using the Huber-White sandwich estimator); bolded values indicate which factor the scale had the largest loading for.

Based on their content and on previous research, we labeled these three factors Pacifism-Aggression, Honesty-Manipulativeness, and Compassion-Callousness. The first factor corresponded to Pacifism-Aggression, having strong positive loadings for the ESI aggression subscales and various relevant scales from the CAT-PD SF and PID-5. The second factor corresponded to Honesty-Manipulativeness, showing the strongest positive loadings for PID-5 Deceitfulness and Manipulativeness, CAT-PD Domineering and Manipulativeness, and Relational Aggression, as well as negative loadings for BFAS Politeness and ESI Honesty. A final factor corresponded to Compassion-Callousness, with strong positive loadings for the BFAS Compassion, IRI empathic concern, and ESI Empathy scales, as well as negative loadings for the CAT-PD SF and PID-5 Callousness scales. Significant cross-loadings were relatively

common, across all factors.

Subsequently, we examined the effects of the Aggression, Callousness, and Manipulativeness factors on ToM accuracy across tasks. Results of the full structural model and ToM measurement model are displayed in Figure 4. (The full measurement model for Agreeableness-Antagonism factors is not displayed here, due to visual complexity). Residual correlations accounting for shared instrument variance are presented in Table A2.

Figure 4.

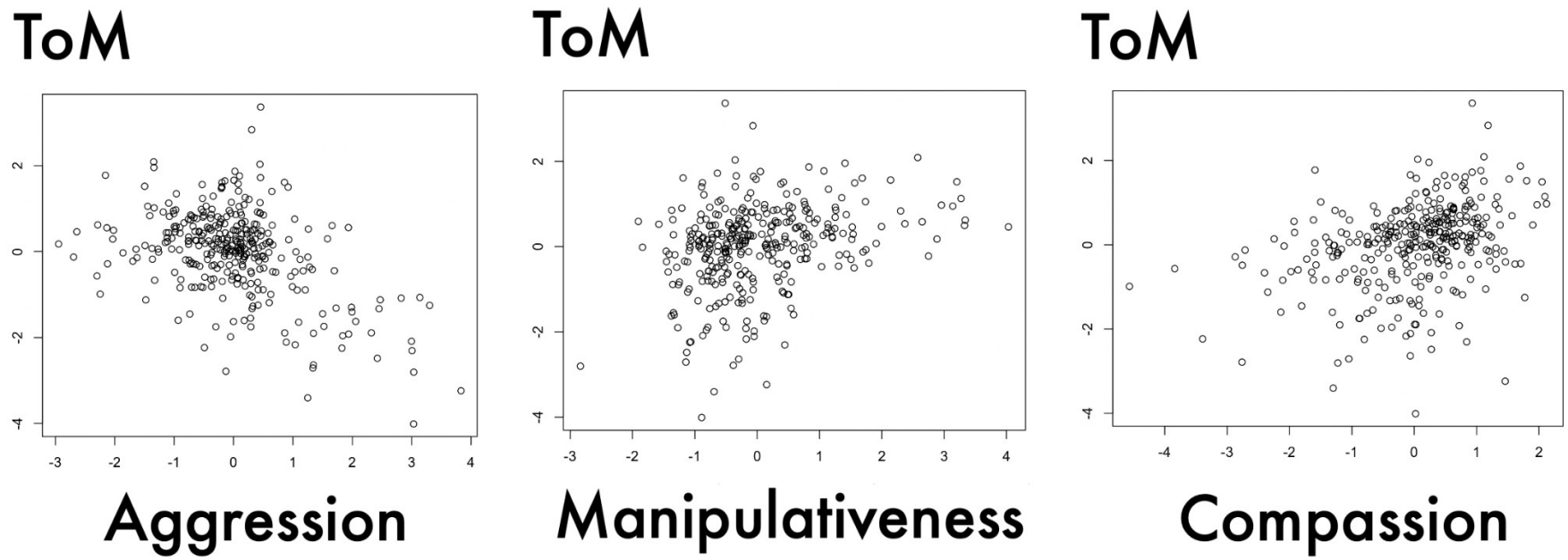


Note. Model of Agreeableness-Antagonism factors and ToM

We found that ToM was negatively predicted by Aggression (95% CI β: [-.96, -.52]) and positively by Compassion (95% CI β: [.08, .41]) as well as Manipulativeness (95% CI β: [.17., .52]). Memory was a positive predictor of ToM (95% CI β: [.12, .46]). Model fit was acceptable (when allowing correlated error terms for scales within each given questionnaire), as indicated by SRMR and RMSEA ≤ .08 and TLI/CFI ≥ .95 (Hooper et al., 2008; Hu & Bentler, 1999). Similar results were obtained whether or not memory was included as a covariate.

Associations among residualized factor score estimates are visualized in Figure 5.

Figure 5.



Note. Scatterplots of ToM and Agreeableness-Antagonism factor score estimates

A final model examined the effects of the Aggression, Manipulativeness, and Compassion factors on observed accuracy scores for the three ToM tasks, controlling for memory performance on the stories task. Factor loadings for Agreeableness-Antagonism scales onto their corresponding factors were nearly equivalent to those obtained in our latent ToM model and are fully reported in the supplement (Table S1). Path coefficients and 95% confidence intervals for Agreeableness-Antagonism factors predicting individual ToM task accuracy scores are presented in Table S2. Effects were all in the same direction of those from the model predicting latent ToM, but their magnitudes were smaller and inconsistently significant; it is possible these individual-task effects were not statistically significant at the $\alpha = .05$ threshold due to sampling variability or the lower reliability of using single tasks as criterion variables.

For both our latent ToM and individual-task models, model specifications that did not freely estimate residual covariances (of subscales taken from the same questionnaire and of memory and ToM accuracy scores from the stories task) yielded results that were substantively equivalent to those reported here. Fit of these models was marginal, as indicated by SRMR $\leq .08$, RMSEA between .08 and .10, and CFI/TLI between .90 and .95 (Hu & Bentler, 1999). Full results from these models are included in our supplement (Figure S1 and Tables S3–S5).

Discussion

The current study investigated how subfactors of the Agreeableness-Antagonism dimension are associated with individual differences in ToM ability. Specifically, we sought to replicate and extend the work of Allen et al. (2017) using multiple behavioral tasks, a more extensive assessment of normal-range and pathological personality traits, and an ESEM approach. Using ESEM, we derived a three-factor structure for a selection of Agreeableness-Antagonism scales. In line with theory and previous work, we labeled these factors Compassion-

Callousness, Pacifism-Aggression, and Honesty-Manipulativeness. Results showed that while higher Compassion and lower Aggression predicted higher ToM ability, higher Manipulativeness also predicted higher ToM ability. These findings replicate the post hoc analyses of Allen et al. (2017) and demonstrate that the effects generalize to multiple behavioral measures of social cognition.

The current findings also extend previous work by Nettle and Liddle (2009), who found that Agreeableness was positively associated with ToM as assessed by the mentalizing vignettes task, but *not* the eyes task. Based on this pattern of results, they concluded that Agreeableness is positively related to "social-cognitive" ToM (understanding mental states based on language and reasoning), but not "social-perceptual" ToM (understanding mental states based on visual or other sensory clues). In contrast, our results indicate that Agreeableness-Antagonism is positively associated with the shared variance *across* ToM tasks, including both the mentalizing vignettes and the eyes task. This suggests that the null finding for Agreeableness and the eyes task reported by Nettle and Liddle may have been due to the use of observed dependent variables, which confound shared variance related to ToM abilities with task-specific unique variance and measurement error. Thus, it seems utilization of latent variable modeling using multiple tasks or multilevel modeling of task data (e.g., Rouder & Haaf, 2019) may be particularly useful when evaluating associations between social cognitive ability and related traits or outcomes. Similar to Allen et al. (2017), we also found that associations between ToM and Agreeableness were specific to certain subfactors; this suggests Nettle and Liddle's finding of an association between ToM and global Agreeableness may have arisen from using a measure highly weighted toward the Compassion aspect. In addition to successfully replicating and

extending previous work, the current study extends our understanding of Agreeableness-Antagonism and has broader implications for personality and individual differences research.

Understanding Agreeableness and Antagonism

It has now been shown in two independent samples that individual differences in social cognitive abilities are positively associated with Compassion and Pacifism but *negatively* related to Honesty, suggesting that better ToM abilities may enable individuals to be more compassionate and less aggressive, but also more deceitful and manipulative. As there is a robust history of research relating Agreeableness—and particularly its Compassion aspect—to questionnaire measures of trait empathy (del Barrio et al., 2004; Graziano et al., 2007; Mooradian et al., 2008; 2011; Penner et al., 1995), it is unsurprising that the compassion subfactor was positively related to ToM ability. The Compassion-Callousness subfactor in the current study showed strong positive loadings for IRI Empathic Concern, BFAS Compassion and ESI Empathy, as well as a strong negative loading for CAT-PD Callousness. Our finding of a positive association between the compassion factor and ToM ability, paired with the previous work of Allen et al. (2017) and Nettle and Liddle (2009), complements research showing that ToM, trait empathy, and Compassion are associated with positive real-world social outcomes (Cassidy et al., 2003; Devine et al., 2016; Stiller & Dunbar, 2007; Sun et al., 2017). Findings are also in line with a long tradition of research documenting social cognitive deficits in those with elevated psychopathy and children with callous-unemotional traits (Jones et al., 2010; Shamay-Tsoory et al., 2010; Pardini et al., 2003). Thus, it is possible that individual differences in the ability to accurately perceive, decipher, and react to the mental states of others is one core psychological mechanism underlying Compassion and prosocial behavior, as well as callousness and antisocial behavior. Future work should further explore the relations between social

cognition and Compassion by incorporating a broader array of behavioral tasks that tap into additional components of social cognitive ability: for instance, tasks capturing additional components of empathy such as emotion contagion or affective resonance (Zaki et al., 2008). It is possible that performance on these tasks would be even more strongly related to Compassion than tasks like those in the current study, which primarily assess cognitive and perceptual aspects of social processing.

In addition to showing a positive association with Compassion, ToM ability was also positively associated with individual differences in Pacifism (or lower scores on the Aggression subfactor). The Pacifism-Aggression subfactor in the current study showed strong positive loadings for several ESI subscales related to aggression, as well as the CAT-PD Hostile Aggression scale. Counterintuitively, the factor loading of PID-5 Callousness was actually higher for the Pacifism-Aggression factor than the Compassion-Callousness factor; this is likely due to the large number of aggression items in that subscale and is consistent with the fact that the current PID-5 Callousness scale was originally designed to measure two separate facets (Callousness and Aggression) but was eventually collapsed into a single scale based on analyses using item response theory (DeYoung et al., 2016; Krueger et al., 2012). Somewhat unexpectedly, the Politeness scale from the BFAS did not significantly load onto the Aggression subfactor (but instead only loaded on the Manipulativeness subfactor). The finding of a negative association between Aggression and ToM result replicates a *post hoc* finding from Allen et al. (2017) and is consistent with other work documenting negative associations between Aggression and ToM ability (Meier et al., 2006; Mohr et al., 2007). These results suggest that facets of Agreeableness other than just Compassion (and trait empathy) are positively associated with individual differences in ToM ability. Future work could better determine to what extent the

specific psychological mechanisms linking lower aggression to better ToM ability are similar to or different from those underlying the association between ToM and Compassion.

The only Agreeableness-Antagonism subfactor showing a diverging pattern of association with ToM ability was Honesty-Manipulativeness. The Manipulativeness subfactor was marked by positive loadings for PID-5 Manipulativeness and Deceitfulness and CAT-PD Manipulativeness and Domineering, as well as negative loadings for BFAS Politeness and ESI Honesty. Paired with the original finding in Allen et al. (2017), the current replication of a positive association between Manipulativeness and ToM provides further empirical support for the longstanding notion that being able to successfully persuade, manipulate, and deceive others is partially reliant on the ability to understand the thoughts and emotions of others (Byrne, 1996; Byrne & Whiten, 1994; Ding et al., 2015; Lonigro et al., 2014; Slaughter et al., 2013; Talwar et al., 2007). Findings are also directly in line with some previous work examining social cognition in the dark triad framework—focusing on the traits of Narcissism, Machiavellianism, and Psychopathy (Furnham et al., 2013; Paulhus & Williams, 2002). Research on these traits and their relation to social cognition is mixed, but tends to suggest that while psychopathy is negatively associated ToM abilities, Machiavellianism and Narcissism (characterized by entitlement, grandiosity, and attention seeking) may be unrelated to or even positively associated with individual differences in ToM (Jonason & Krause, 2013; Kajonius & Björkman, 2020; Paal & Bereczkei, 2007; Schimmenti et al., 2019; Stellwagen & Kerig, 2013; Vonk et al., 2015; Wai & Tiliopoulos, 2012). Taken with previous work and theory, the current findings seem to suggest that specific traits often conceptualized as pathological (i.e., dishonesty and manipulativeness) may actually be associated with enhanced abilities and outcomes in some contexts. Indeed, this is consistent with how many researchers have discussed potential adaptive qualities of

Machiavellianism, which can be particularly apparent in high-power occupations such as attorneys, executives, and salespeople (Byrne & Whiten, 1994; Furnham & Treglown, 2021; Grover & Furnham, 2021). Nonetheless, whether being able to successfully deceive others leads to positive functional outcomes at the individual and group level will typically depend on situational context and one's other trait levels. Indeed, those with the highest levels of ToM ability in the current dataset have both high Manipulativeness and Compassion, suggesting these individuals may use deceit and manipulation for prosocial means.

It is worth noting that the current results contrast with a few studies that found a negative association or no association between ToM ability and Machiavellianism (e.g., Ali & Chamorro-Premuzic, 2010; Lyons et al., 2010). To explain these associations, some have suggested those with elevated dark triad traits possess the capacity and ability to engage in ToM-related processes, but lack the disposition—and, in most situations, the motivation—to do so (Kajonius & Björkman, 2020). Nonetheless, scales used to assess Machiavellianism in many of these studies from the dark triad perspective tend to conflate dishonesty and manipulation with other facets of Antagonism, such as immorality or mistrust, indicating that the Honesty-Manipulativeness dimension as construed in the current work and by Allen at al. (2017) may, indeed, be associated with patterns of ToM that diverge from associations with the broader Agreeableness-Antagonism domain. Moreover, studies that do not control for subfactors predicting ToM abilities in opposite directions may have true effects in either direction masked due to statistical suppression (Martinez Gutierrez & Cribbie, 2021; Tzelgov & Henik, 1991).

Additional Implications for Individual Differences Research

Individual differences in social cognition are conceptually similar to several other constructs that exist in disparate lines of work—for instance emotional intelligence and

Gardner's interpersonal intelligence (Gardner, 2011; McEnrue & Groves, 2006). In recent years, emotional intelligence has gained widespread popularity as a construct in both popular culture and among researchers, with some arguing emotional intelligence is predictive of broad positive life outcomes and should be used to inform hiring decisions (Bar-On, 2001; Emanuel & Gudbranson, 2018; Fox & Spector, 2000; Goleman, 1996; Stein & Book, 2011; Watkin, 2000; van der Linden et al., 2010; 2012; 2017). In contrast, others have argued that measures of emotional intelligence (particularly those relying only on self-report) provide little incremental validity over general intelligence and Conscientiousness, when it comes to predicting important occupational and educational outcomes (Amelang & Steinmayr, 2006; Antonakis, 2004; Gottfredson, 1997; Landy, 2005; Ones et al., 2012; Schmidt et al., 2008; Thorndike & Stein, 1937; Van Rooy & Viswesvaran, 2004; Willoughby & Boutwell, 2018). Our current results add to a growing body of research suggesting that while performance-based measures of social cognition may be highly correlated with measures of general cognitive ability, they do indeed show incremental validity in predicting socially relevant personality traits and outcomes (Allen et al., 2017; Stiller & Dunbar, 2007).

The current study provides insights into the theory and measurement of social cognition and its association with Agreeableness-Antagonism subfactors. Because problems with ToM and related interpersonal outcomes are characteristic of multiple mental disorders and symptom domains, elucidating their association with normal-range personality traits and improving their measurement may eventually help facilitate more effective methods for assessment and treatment. Such an approach is in line with the Hierarchical Taxonomy of Psychopathology's conceptualization of psychiatric illness, the National Institute of Mental Health's (NIMH's) Research Domain Criteria (RDoC) initiative, and theories of psychopathology that emphasize

continuity with normal personality variation and impairments in cybernetic functioning (DeYoung & Krueger, 2018; Insel et al., 2010; Kotov et al., 2017). Future work on this topic could incorporate additional tasks to span a range of social cognitive and interpersonal abilities, including more of those recommended by the NIMH's Workgroup on Tasks and Measures for RDoC (Barch et al., 2016). Furthermore, it remains to be seen whether the personality correlates of social cognitive abilities in populations with more extreme levels of Antagonism (i.e., criminal offenders or those diagnosed with antisocial or narcissistic personality disorder) would reflect the patterns observed in the general population. Finally, another topic worth exploring further is whether the personality correlates and mechanisms of social cognitive deficits are consistent or divergent across different psychopathology dimensions; in particular, research should explore whether ToM deficits associated with psychoticism and positive schizotypy or autistic traits and symptoms are similar in etiology and mechanisms to those associated with Antagonism and related personality disorders.

Limitations

Though the current study had numerous strengths, several limitations are worth noting. First, the current sample had an overrepresentation of females and people of European and Asian ancestry; future work should attempt to collect more demographically representative samples. The current study's measures of Agreeableness-Antagonism were self-reported and could be usefully supplemented in future research by peer reports or clinician ratings. Also, this work would have further benefited from the inclusion of a general intelligence measure such as the Wechsler Adult Intelligence Scale or International Cognitive Ability Resource (Condon & Revelle, 2014; Wechsler, 2008), as including such measures would allow us to better parse associations between personality and social cognition without the influence of general cognitive

ability; such measures could also allow researchers to directly test how individual differences in general cognitive ability might be associated with Agreeableness-Antagonism and its constituent subfactors. These limitations could be usefully addressed by recruiting additional large samples with extensive, high-quality measures of personality, social cognition, general cognition, and real-world social functioning including peer-report and experience sampling data.

Additionally, although the current set of questionnaires captured a broad range of Agreeableness-Antagonism facets, future research could also incorporate measures from the HEXACO and Dark Triad literatures (Collison et al., 2018; Jonason & Webster, 2010; Jones & Paulhus, 2014; Lee & Ashton, 2004; Ashton & Lee, 2007). This approach could directly test whether the constructs represented in those measures (i.e., Honesty-Humility and Machiavellianism) map onto the Honesty-Manipulativeness dimension revealed in the current study using measures based on the Five Factor Model, and whether Honesty-Humility and Machiavellianism show corresponding associations with ToM. Future work could also intentionally recruit participants with clinically significant levels of Antagonism and other traits related to social cognitive deficits; this could be a particularly fruitful line of work if the tools used to assess social cognitive ability and personality domains such as Agreeableness could eventually be used to foster more effective early detection and intervention for forms of psychopathology related to interpersonal dysfunction.

Conclusion

Agreeableness-Antagonism is robustly related to life outcomes, including victimization, relationship satisfaction, aggression, and a variety of psychiatric disorders (Gore & Widiger, 2013; Lynam & Miller, 2019). Despite its enormous consequences, however, Agreeableness-Antagonism is arguably the least studied dimension of the Big Five and their pathological

counterparts (Gore & Widiger, 2013; Lynam & Miller, 2019). The current research improves the scientific understanding of Agreeableness-Antagonism, replicating and extending work that suggests differential relations of Agreeableness-Antagonism subfactors with social cognition. Our findings suggest ToM abilities might facilitate individual differences in most traits related to Agreeableness, with a distinctly negative association with specific honesty-related tendencies. This paradox adds to a set of interesting similar patterns where the correlates or outcomes of personality traits diverge at levels below the Big Five, further underscoring the importance of facet-level research and parsing the subfactors of broad personality domains. Future research should more thoroughly examine situational, motivational, and relationship-specific factors to determine their potential moderating role in the associations between personality and social cognitive abilities.

References

Allen, T. A., Rueter, A. R., Abram, S. V., Brown, J. S., & DeYoung, C. G. (2017). Personality and neural correlates of mentalizing ability. *European Journal of Personality, 31,* 599–613.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (5th ed.).* Washington, DC: American Psychiatric Association.

Barch, D. M., Oquendo, M. A., Pacheco, J., & Morris, S. (2016). *Behavioral assessment methods for RDoC constructs: A report by the National Advisory Mental Health Council Workgroup on tasks and measures for RDoC.* Washington, DC: National Institutes of Mental Health.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioral and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology, 4,* 113–125.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science, 20,* 286–290.

Blain, S. D., Grazioplene, R. G., Ma, Y., & DeYoung, C. G. (2020). Toward a neural model of the Openness-Psychoticism dimension: Functional connectivity in the default and frontoparietal control networks. *Schizophrenia Bulletin, 46*(3), 540-551.

Blain, S. D., Longenecker, J. M., Grazioplene, R. G., Klimes-Dougan, B., & DeYoung, C. G. (2020). Apophenia as the disposition to false positives: A unifying framework for openness and psychoticism. *Journal of Abnormal Psychology, 129*(3), 279.

Blain, S. D., Sassenberg, T. A., Xi, M., Zhao, D., & DeYoung, C. G. (2020). Extraversion but not depression predicts reward sensitivity: Revisiting the measurement of anhedonic

phenotypes. *Journal of Personality and Social Psychology.*

Bora, E., & Pantelis, C. (2013). Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. *Schizophrenia Research, 144*(1-3), 31-36.

Byrne, R., & Whiten, A. (1994). *Machiavellian Intelligence.* Oxford: Oxford University Press.

Cassidy, K. W., Werner, R. S., Rourke, M., Zubernis, L. S., & Balaraman, G. (2003). The relationship between psychological understanding and positive social behaviors. *Social Development, 12*(2), 198–221.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43,* 52–64.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10,* 85.

del Barrio, V., Aluja, A., & García, L. F. (2004). Relationship between empathy and the Big Five personality traits in a sample of Spanish adolescents. *Social Behavior and Personality: An International Journal, 32*(7), 677–681.

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

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of research in personality, 56,* 33-58.

DeYoung, C. G. & Blain, S. D. (2020). Personality neuroscience: Explaining individual

differences in motivation, emotion, cognition, and behavior. In Corr, P. J., & Matthews, G. (Eds.) *The Cambridge Handbook of Personality Psychology* (2nd ed.). New York: Cambridge University Press.

DeYoung, C. G., Grazioplene, R. G., & Peterson, J. B. (2012). From madness to genius: The Openness/Intellect trait domain as a paradoxical simplex. Journal of Research in Personality, 46(1), 63-78.

DeYoung, C. G., Carey, B. E., Krueger, R. F., & Ross, S. R. (2016). Ten aspects of the Big Five in the Personality Inventory for DSM–5. *Personality Disorders: Theory, Research, and Treatment, 7*(2), 113–123.

DeYoung, C. G., & Krueger, R. F. (2018). A cybernetic theory of psychopathology. *Psychological Inquiry, 29*(3), 117–138.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93,* 880–896.

DeYoung, C. G., Quilty, L. C., Peterson, J. B., & Gray, J. R. (2014). Openness to experience, intellect, and cognitive ability. *Journal of personality assessment, 96*(1), 46-52.

Dolan, M. & Fullam, R. (2004). Theory of mind and mentalizing ability in antisocial personality disorders with and without psychopathy. *Psychological Medicine, 34,* 1093–1102.

Furnham, A., & Treglown, L. (2021). The dark side of high-fliers: The Dark Triad, high-flier traits, engagement, and subjective success. *Frontiers in Psychology, 12.* https://doi.org/%E2%80%8B10.3389/fpsyg.2021.647676

Gardner, H. (2011). *Frames of Mind: The Theory of Multiple Intelligences.* Hachette UK.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower levels of several five-factor models. In F. Ostendorf (Ed.), *Personality psychology*

in Europe (Volume 7, pp. 7–28). Tilburg: Tilburg University Press.

Gore, W. L., & Widiger, T. A. (2013). The DSM-5 dimensional trait model and five-factor models of general personality. *Journal of Abnormal Psychology, 122,* 816–821.

Graziano, W. G., Habashi, M. M., Sheese, B. E., & Tobin, R. M. (2007). Agreeableness, empathy, and helping: A person × situation perspective. *Journal of Personality and Social Psychology, 93*(4), 583–599.

Grover, S., & Furnham, A. (2021). The Dark Triad, emotional intelligence, self-monitoring and executive coach effectiveness and satisfaction. *Coaching: An International Journal of Theory, Research and Practice,* 1–21.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Evaluating model fit: a synthesis of the structural equation modelling literature. In the 7th *European Conference on research methodology for business and management studies*,195-200.

Hopwood, C. J., Thomas, K. M., Markon, K. E., Wright, A. G., & Krueger, R. F. (2012). DSM-5 personality traits and DSM-IV personality disorders. *Journal of Abnormal Psychology, 121,* 424–432.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry, 167*(7), 748–751.

Jensen-Campbell, L. A., Adams, R., Perry, D. G., Workman, K. A., Furdella, J. Q., & Egan, S. K. (2002). Agreeableness, extraversion, and peer relations in early adolescence: Winning

friends and deflecting aggression. *Journal of Research in Personality, 36,* 224–251.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.). *Handbook of Personality: Theory and Research* (p. 114–158). New York: Guilford.

Jonason, P. K., & Krause, L. (2013). The emotional deficits associated with the dark triad traits: Cognitive empathy, affective empathy, and alexithymia. *Personality and Individual Differences, 55*(5), 532–537.

Kajonius, P. J., & Björkman, T. (2020). Individuals with dark traits have the ability but not the disposition to empathize. *Personality and Individual Differences, 155,* 109716.

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*(5), 768–821.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., ... & Eaton, N. R. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126*(4), 454–477.

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine, 42,* 1879–1890.

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology, 116,* 645–666.

Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology, 4,* 231–244.

Lonigro, A., Laghi, F., Baiocco, R., & Baumgartner, E. (2014). Mind reading skills and empathy: Evidence for nice and nasty ToM behaviours in school-aged children. *Journal of Child and Family Studies, 23*(3), 581–590.

Lucas, R. E., Diener, E., Grob, A., Suh, E. M., & Shao, L. (2000). Cross-cultural evidence for the fundamental features of extraversion. Journal of Personality and Social Psychology, 79(3), 452–468. https://doi.org/10.1037/0022-3514.79.3.452

Lyons, M., Caldwell, T., & Shultz, S. (2010). Mind-reading and manipulation—Is Machiavellianism related to theory of mind? *Journal of Evolutionary Psychology, 8,* 261–274.

Martinez Gutierrez, N., & Cribbie, R. (2021). Incidence and interpretation of statistical suppression in psychological research. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement.*

McDonald, R. P. (1999). *Test theory: A unified treatment.* Lawrence Earlbaum Associates. Inc., Mahwah, NJ, 142–145.

McEnrue, M. P., & Groves, K. (2006). Choosing among tests of emotional intelligence: what is the evidence? *Human Resource Development Quarterly, 17*(1), 9–42.

Meier, B. P., Robinson, M. D., & Wilkowski, B. M. (2006). Turning the other cheek: Agreeableness and the regulation of aggression-related primes. *Psychological Science, 17,* 136–142.

Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar personality patterns are associated with empathy in four different countries. *Frontiers in*

Psychology, 7, 1–12.

Mohr, P., Howells, K., Gerace, A., Day, A., & Wharton, M. (2007). The role of perspective taking in anger arousal. *Personality and Individual Differences, 43,* 507–517.

Mooradian, T. A., Davis, M., & Matzler, K. (2011). Dispositional empathy and the hierarchical structure of personality. *The American Journal of Psychology, 124*(1), 99–109.

Mooradian, T. A., Matzler, K., & Szykman, L. (2008). Empathetic responses to advertising: Testing a network of antecedents and consequences. *Marketing Letters, 19*(2), 79–92.

Muthén, L.K. and Muthén, B.O. (2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Nettle, D., & Liddle, B. (2008). Agreeableness is related to social-cognitive, but not social-perceptual, theory of mind. *European Journal of Personality, 22,* 323–335.

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences, 43,* 541–551.

Patrick, C. J., Kramer, M. D., Krueger, R. F., & Markon, K. E. (2013). Optimizing efficiency of psychopathology assessment through quantitative modeling: Development of a brief form of the Externalizing Spectrum Inventory. *Psychological Assessment, 25,* 1332–1348.

Penner, L. A., Fritzsche, B. A., Craiger, J. P., Freifeld, T. R., Butcher, J. N., & Spielberger, C. D. (1995). Measuring the prosocial personality. *Advances in Personality Assessment, 10,* 147–163.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1,* 515–526.

Preißler, S., Dziobek, I., Ritter, K., Heekeren, H. R., & Roepke, S. (2010). Social cognition in

borderline personality disorder: evidence for disturbed recognition of the emotions, thoughts, and intentions of others. *Frontiers in Behavioral Neuroscience, 4,* 182.

Quilty, L. C., Ayearst, L., Chmielewski, M., Pollock, B. G., & Bagby, R. M. (2013). The psychometric properties of the personality inventory for DSM-5 in an APA DSM-5 field trial sample. *Assessment, 20,* 362–369.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment, 31*(12), 1395–1411.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review, 26*(2), 452-467.

Schimmenti, A., Jonason, P. K., Passanisi, A., La Marca, L., Di Dio, N., & Gervasi, A. M. (2019). Exploring the dark side of personality: Emotional awareness, empathy, and the Dark Triad traits in an Italian Sample. *Current Psychology, 38*(1), 100–109.

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT-PD Project. *Journal of Personality Assessment, 93,* 380–389.

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current Biology, 24,* R875–R878.

Slaughter, V., Peterson, C. C., & Moore, C. (2013). I can talk you into it: Theory of mind and persuasion behavior in young children. *Developmental Psychology, 49*(2), 227–231.

Smillie, L. D., Cooper, A. J., Wilt, J., & Revelle, W. (2012). Do extraverts get more bang for the buck? Refining the affective-reactivity hypothesis of Extraversion. *Journal of Personality and Social Psychology, 103*(2), 306–326.

Smillie, L. D., Jach, H. K., Hughes, D. M., Wacker, J., Cooper, A. J., & Pickering, A. D. (2019).

Extraversion and reward-processing: Consolidating evidence from an electroencephalographic index of reward-prediction-error. *Biological psychology, 146,* 107735.

Stellwagen, K. K., & Kerig, P. K. (2013). Dark triad personality traits and theory of mind among school-age children. *Personality and Individual Differences*, *54*(1), 123–127.

Sun, J., Kaufman, S. B., & Smillie, L. D. (2018). Unique associations between Big Five personality aspects and multiple dimensions of well-being. *Journal of Personality, 86*(2), 158–172.

Suzuki, T., Samuel, D. B., Pahlen, S., & Krueger, R. F. (2015). DSM-5 alternative personality disorder model traits as maladaptive extreme variants of the five-factor model: An item-response theory analysis. *Journal of Abnormal Psychology, 124*(2), 343–354.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological bulletin, 109*(3), 524.

Venables, N. C., & Patrick, C. J. (2012). Validity of the Externalizing Spectrum Inventory in a criminal offender sample: Relations with disinhibitory psychopathology, personality, and psychopathic features. *Psychological Assessment, 24,* 88–100.

Vonk, J., Zeigler-Hill, V., Ewing, D., Mercer, S., & Noser, A. E. (2015). Mindreading in the dark: Dark personality features and theory of mind. *Personality and Individual Differences, 87,* 50–54.

Wai, M., & Tiliopoulos, N. (2012). The affective and cognitive empathic nature of the dark triad of personality. *Personality and Individual Differences, 52*(7), 794–799.

Wechsler, D., Psychological Corporation., & Pearson Education, Inc. (2008). *WAIS-IV: Wechsler adult intelligence scale.*

Wright, A. G. C., Pincus, A. L., Hopwood, C. J., Thomas, K. M., Markon, K. E., & Krueger, R. F. (2012). An interpersonal analysis of pathological personality traits in DSM-5. *Assessment, 19,* 263–275.

Wright, A. G. C., & Simms, L. J. (2014). On the structure of personality disorder traits: Conjoint analysis of the CT-PD, PID-5, and NEO PI-3 trait models. *Personality Disorders: Theory, Research, and Treatment, 5,* 43–54.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19*(4), 399–404.

Appendix

Table A1. Bivariate correlations of self-report measures and task performance

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.
1. Eyes Task	1.00																				
2. Triangles Task	.47	1.00																			
3. Dunbar ToM	.50	.37	1.00																		
4. Dunbar Mem	.53	.40	.72	1.00																	
5. BFAS Compassion	.31	.21	.28	.23	1.00																
6. BFAS Politeness	.23	.19	.21	.22	.58	1.00															
7. IRI Empathic Concern	.35	.30	.28	.27	.63	.49	1.00														
8. PID-5 Callousness	-.51	-.41	-.43	-.49	-.60	-.57	-.58	1.00													
9. PID-5 Manipulativeness	-.32	-.24	-.21	-.35	-.40	-.53	-.40	.67	1.00												
10. PID-5 Deceitfulness	-.36	-.31	-.28	-.38	-.54	-.60	-.50	.78	.80	1.00											
11. CAT-PD Callousness	-.49	-.37	-.39	-.44	-.70	-.58	-.66	.81	.60	.71	1.00										
12. CAT-PD Domineering	-.30	-.25	-.24	-.33	-.52	-.65	-.47	.66	.69	.70	.73	1.00									
13. CAT-PD Hostile Aggression	-.55	-.44	-.47	-.56	-.54	-.59	-.52	.84	.61	.70	.83	.72	1.00								
14. CAT-PD Manipulativeness	-.44	-.40	-.38	-.47	-.55	-.64	-.57	.78	.70	.81	.82	.78	.85	1.00							
15. ESI Theft	-.49	-.41	-.43	-.53	-.39	-.42	-.40	.71	.57	.65	.65	.59	.77	.81	1.00						
16. ESI Fraud	-.51	-.40	-.42	-.54	-.46	-.47	-.44	.76	.62	.72	.69	.63	.77	.76	.87	1.00					
17. ESI Honesty	.17	.26	.18	.21	.27	.33	.47	-.32	-.33	-.46	-.36	-.31	-.34	-.46	-.26	-.29	1.00				
18. ESI Physical Aggression	-.49	-.40	-.41	-.52	-.45	-.45	-.43	.75	.58	.64	.67	.61	.81	.72	.85	.84	-.20	1.00			
19. ESI Destructive Aggression	-.56	-.45	-.46	-.56	-.46	-.46	-.41	.79	.56	.65	.69	.60	.82	.74	.88	.88	-.21	.89	1.00		
20. ESI Relational Aggression	-.43	-.36	-.32	-.42	-.51	-.61	-.48	.77	.67	.76	.75	.71	.79	.81	.80	.85	-.31	.83	.83	1.00	
21. ESI Empathy	.48	.42	.41	.43	.73	.55	.74	-.74	-.53	-.63	-.80	-.63	-.71	-.71	-.61	-,64	.44	-.66	-.65	-.69	1.00

Notes. $N = 335$. Correlations of $r \geq .11$ are significant (at an α of .05). BFAS = Big Five Aspect Scales, IRI = Interpersonal Reactivity Inventory, CAT = Computer Adaptive Test of Personality Disorders Static Form, PID-5 = Personality Inventory for DSM-5, ESI = Externalizing Spectrum Inventory.

Table A2. Residual correlations accounting for shared instrument variance in models of Agreeableness-Antagonism factors and ToM

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.
1. Dunbar ToM		—																	
2. Dunbar Mem	.39*																		
3. BFAS Compassion				.24*															
4. BFAS Politeness			.24*																
5. IRI Empathic Concern																			
6. PID-5 Callousness							.29*	.38*											
7. PID-5 Manipulativeness						.29*		.39*											
8. PID-5 Deceitfulness						.37*	.39*												
9. CAT-PD Callousness										.22*	.34*	.29*							
10. CAT-PD Domineering									.22*		.33*	.16							
11. CAT-PD Hostile Aggression									.34*	.33*		.30*							
12. CAT-PD Manipulativeness									.29*	.16	.30*								
13. ESI Theft														.55*	.11	.47*	.56*	.37*	-.21
14. ESI Fraud													.55*		-.13	.44*	.57*	.44*	-.11
15. ESI Honesty													.11	.13		.25*	.27*	.23*	.21*
16. ESI Physical Aggression													.47*	.44*	.25*		.54*	.49*	-.22*
17. ESI Destructive Aggression													.56*	.57*	.27*	.54*		.47*	-.10
18. ESI Relational Aggression													.37*	.44*	.23*	.49*	.47*		-.16
19. ESI Empathy													-.21	-.11	.21*	-.22*	-.10	-.16	

Notes. $N = 335$. *$p < .05$. Residual correlations from Model 2 (latent ToM) are displayed below the diagonal and those from Model 3 (individual tasks) are displayed above. BFAS = Big Five Aspect Scales, IRI = Interpersonal Reactivity Inventory, CAT = Computer Adaptive Test of Personality Disorders Static Form, PID-5 = Personality Inventory for DSM-5, ESI = Externalizing Spectrum Inventory.

Supplemental Methods and Results

Table S1. Factor loadings of Agreeableness-Antagonism scales on three exploratory factors for the individual ToM task model

Scale	Aggression	Manipulativeness	Compassion
CAT – Domineering	.21	**.57***	-.18*
CAT – Manipulativeness	**.47***	.46*	-.14*
PID – Manipulativeness	.23	**.61***	-.02
PID – Deceitfulness	.27	**.61***	-.12*
ESI – Honesty	-.02	**-.39***	.15
BFAS – Politeness	-.07	**-.51***	.26*
ESI – Relational Aggression	**.53***	.39*	-.10*
ESI – Physical Aggression	**.82***	.03	-.03
ESI – Destructive Aggression	**.89***	-.02	-.03
ESI – Theft	**.78***	.13	.05
ESI – Fraud	**.70***	.20	-.03
CAT – Hostile Aggression	**.82***	.05	-.14*
PID – Callousness	**.66***	.06	-.31*
CAT – Callousness	.40*	.15*	**-.52***
ESI – Empathy	-.29*	-.05	**.69***
BFAS – Compassion	-.02	-.08	**.78***
IRI – Empathic Concern	-.05	-.10	**.71***

Note. *$p < .05$ (based on the z-distribution and standard errors computed using the Huber-White sandwich estimator)

Table S2. Model predicting individual ToM task accuracy from Agreeableness-Antagonism factors

Criterion Variable	*β*	95% CI	*p*
Dunbar ToM Stories			
Aggression	-.15	[-.29, -.01]	.036
Manipulativeness	.19	[.07, .31]	.002
Compassion	.16	[.05, .26]	.003
Memory	.64	[.53, .74]	< .001
Eyes Task			
Aggression	-.49	[-.68, -.30]	< .001
Manipulativeness	.23	[.09, .38]	.002
Compassion	.19	[.06, .32]	.005
Memory	.24	[.10, .37]	.001
Triangles Task			
Aggression	-.42	[-.65, -.20]	< .001
Manipulativeness	.13	[-.03, .30]	.114
Compassion	.11	[-.03, .25]	.126
Memory	.15	[.00, .29]	.045

Table S3. Factor loadings of Agreeableness-Antagonism scales on three exploratory factors for the latent ToM model (restricting residual covariances)

Scale	Aggression	Manipulativeness	Compassion
CAT – Domineering	.20*	**.56***	-.21*
CAT – Manipulativeness	.37*	**.48***	-.25*
PID – Manipulativeness	.17*	**.76***	.04
PID – Deceitfulness	.22*	**.71***	-.10*
ESI – Honesty	.14	**-.36***	.32*
BFAS – Politeness	-.02	**-.48***	.34*
CAT – Hostile Aggression	**.59***	.19*	-.29*
ESI – Physical Aggression	**.88***	.02	-.07*
ESI – Relational Aggression	**.63***	.32*	-.10*
ESI – Destructive Aggression	**.95***	-.03	-.06*
ESI – Theft	**.88***	.08	.02
ESI – Fraud	**.81***	.17*	-.01
PID – Callousness	**.48***	.25*	-.34*
CAT – Callousness	.30*	.22*	**-.57***
ESI – Empathy	-.29*	-.02	**.72***
BFAS – Compassion	-.04	-.07	**.74***
IRI – Empathic Concern	-.01	-.05	**.77***

Note. *$p < .05$ (based on the z-distribution and standard errors computed using the Huber-White sandwich estimator)

Figure S1. Model of Agreeableness-Antagonism factors and ToM (restricting residual covariances)

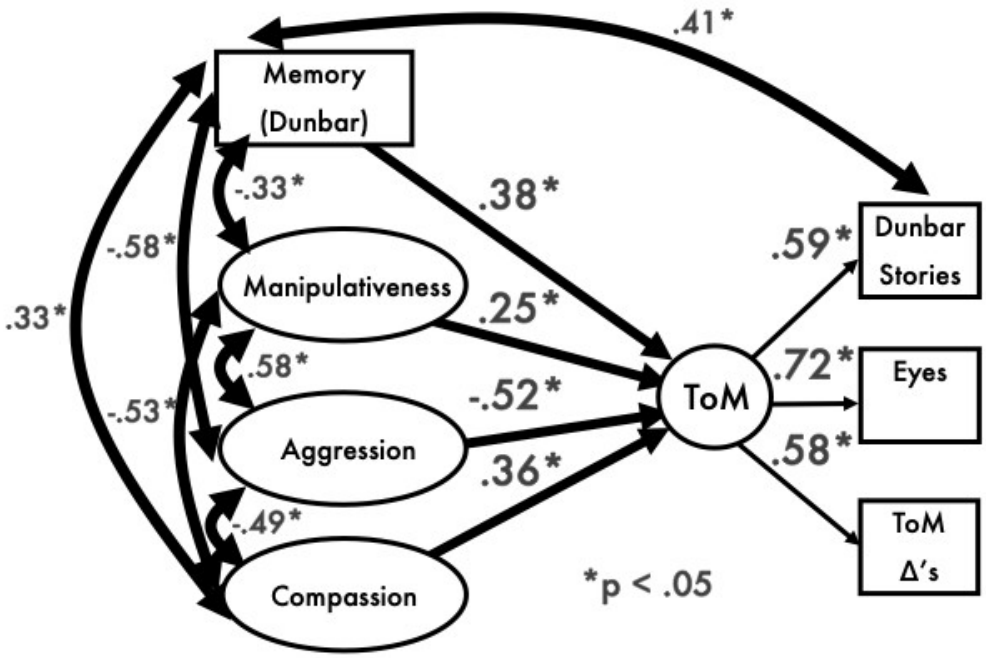


Table S4. Factor loadings of Agreeableness-Antagonism scales on three exploratory factors for the individual ToM task model (in model restricting residual covariances)

Scale	Aggression	Manipulativeness	Compassion
CAT – Domineering	.20*	**.56***	-.21*
CAT – Manipulativeness	.37*	**.48***	-.25*
PID – Manipulativeness	.17*	**.76***	.04
PID – Deceitfulness	.22*	**.71***	-.10*
ESI – Honesty	.14*	**-.36***	.32*
BFAS – Politeness	-.02	**-.47***	.34*
CAT – Hostile Aggression	**.59***	.18*	-.30*
ESI – Physical Aggression	**.88***	.02	-.07*
ESI – Relational Aggression	**.63***	.32*	-.10*
ESI – Destructive Aggression	**.95***	-.03	-.06*
ESI – Theft	**.88***	.08	.02
ESI – Fraud	**.81***	.17*	-.01
PID – Callousness	**.48***	.25*	-.34*
CAT – Callousness	.30*	.22*	**-.58***
ESI – Empathy	-.29*	-.02	**.71***
BFAS – Compassion	-.04	-.07	**.75***
IRI – Empathic Concern	.00	-.05	**.77***

Note. *$p < .05$ (based on the z-distribution and standard errors computed using the Huber-White sandwich estimator)

Table S5. Model predicting individual ToM task accuracy from Agreeableness-Antagonism factors (restricting residual covariances)

Criterion Variable	*β*	95% CI	*p*
Dunbar ToM Stories			
Aggression	-.08	[-.21, .05]	.234
Manipulativeness	.17	[.05, .28]	.004
Compassion	.20	[.10, .30]	< .001
Memory	.66	[.57, .76]	< .001
Eyes Task			
Aggression	-.36	[-.52, -.21]	< .001
Manipulativeness	.16	[.02, .30]	.027
Compassion	.24	[.11, .36]	< .001
Memory	.29	[.17, .41]	< .001
Triangles Task			
Aggression	-.29	[-.46, -.13]	.001
Manipulativeness	.10	[-.04, .24]	.172
Compassion	.20	[.06, .34]	.005
Memory	.20	[.07, .33]	.002