Die Dunkle Triade und Täuschung

Dissertation zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.)

Vorgelegt im Fachbereich Humanwissenschaften

der Universität Kassel

von Benno Gerrit Wissing

Kassel, Februar 2021

(Datum der Disputation: 27.10.2021)

Betreuer:

Prof. Dr. Marc-André Reinhard

Dr. Simon Schindler

Inhaltsverzeichnis

Übersicht 4

Synopse 6

I. Biologisch evolvierte Täuschung 7

Die Dunkle Triade und Täuschung 8

Die Dunkle Triade 8

Die Evolution von Persönlichkeitseigenschaften und Täuschung 9

Täuschungsentdeckungsfähigkeit 10

Täuschungsproduktionsfähigkeit 11

Generelle Täuschungsfähigkeit 14

Selbstwahrgenommene Täuschungsproduktionsfähigkeit 15

Diskussion 16

The Dark Triad and Deception Perceptions 16

The Dark Triad and the PID-5 Maladaptive Personality Traits: Accuracy, Confidence and Response Bias in Judgments of Veracity 19

Konklusion 21

II. Technologisch evolvierte Täuschung 23

Künstliche Intelligenz, Persönlichkeitseigenschaften und Täuschung 24

Existenzielles Risiko 25

Sicherheitsforschung 26

Täuschung 27

Täuschungsmotivation 29

Menschliche Faktoren 30

Adaptivität der Dunklen Triade 32

Diskussion 34

Individual Differences in Risk Perception of Artificial Intelligence 34

Konklusion 36

Literaturverzeichnis 38

Erklärung Eigenanteil 51

Anhang 54

Übersicht

Die vorliegende Dissertationsschrift gliedert sich in zwei Teile: die Synopse und einen empirischen Teil (Anhänge A bis C). In der Synopse werden die empirischen Ergebnisse zusammengefasst und mit bestehenden Forschungsergebnissen kontextualisiert. Der empirische Teil umfasst alle Studien, die im Rahmen der Dissertation durchgeführt und publiziert wurden (Wissing & Reinhard, 2017, 2018, 2019).

Die Studien liefern Daten zu dem Zusammenhang zwischen den Persönlichkeitseigenschaften der sogenannten Dunklen Triade (*Dark Triad*; DT; Paulhus & Williams, 2002) – bestehend aus den moderat interkorrelierten Persönlichkeitseigenschaften Narzissmus, Machiavellismus und Psychopathie – und multiplen Täuschungsvariablen, insbesondere Täuschungsentdeckungsfähigkeit, Antworttendenz (*truth bias*), selbstwahrgenommene Täuschungsentdeckungsfähigkeit, selbstwahrgenommene Täuschungsproduktionsfähigkeit und selbstwahrgenommene Täuschungsentdeckbarkeit.

Die erste Studie (Anhang A; Wissing & Reinhard, 2019) untersucht den Zusammenhang der DT-Persönlichkeitseigenschaften mit selbstwahrgenommener Täuschungsproduktionsfähigkeit, selbstwahrgenommener Täuschungsentdeckungsfähigkeit und selbstwahrgenommener Täuschungsentdeckbarkeit. Die zweite Studie (Anhang B; Wissing & Reinhard, 2017) untersucht den Zusammenhang der DT-Persönlichkeitseigenschaften mit Täuschungsentdeckungsfähigkeit, selbstwahrgenommener Täuschungsentdeckungsfähigkeit und Antworttendenz (*truth bias*). Die dritte Studie (Anhang C; Wissing & Reinhard, 2018) untersucht den Zusammenhang der DT-

Persönlichkeitseigenschaften mit Risikowahrnehmung von künstlicher Intelligenz (KI), insbesondere von Risikoszenarien, in denen KIs Täuschung anwenden, um ihre Ziele zu erreichen (Bostrom, 2014).

Die Ergebnisse der Studien konvergieren dabei auf ein Assoziationsmuster der DT-Persönlichkeitseigenschaften mit Täuschungsvariablen, das von substanziellen Verbindungen zu selbstwahrgenommenen Täuschungsfähigkeiten und der Abwesenheit von Verbindungen zu tatsächlichen Täuschungsfähigkeiten geprägt ist.

Synopse

I. Biologisch evolvierte Täuschung

Die Dunkle Triade und Täuschung

Im Folgenden wird die Literatur zum Zusammenhang der DT-Persönlichkeitseigenschaften mit zentralen Täuschungsvariablen zusammengefasst.

Die Dunkle Triade

Persönlichkeitseigenschaften sind ein entscheidender und robuster Prädiktor von Lebensergebnissen über ökonomische und soziologische Variablen hinaus (z.B. Beck, 2020). Unter den Persönlichkeitseigenschaften hat die Forschung zu jenen der DT (Paulhus & Williams, 2002) – konstituiert aus den mäßig korrelierten Persönlichkeitseigenschaften von Narzissmus, Machiavellismus und Psychopathie – eine umfassende Literatur produziert. Innerhalb der beiden etablierten Persönlichkeitstheorieräume, der *Big Five* (Costa & McCrea, 1992) und des HEXACO-Modells (Lee & Ashton, 2005), konvergieren die DT-Persönlichkeitseigenschaften primär auf geringe Ausprägungen von Verträglichkeit (Paulhus & Williams, 2002) und *Honesty-Humility* (Book, Visser, & Volk, 2015; Lee & Ashton, 2005). Auf der zwischenmenschlichen Ebene sind Personen mit hohen Ausprägungen der DT-Persönlichkeitseigenschaften antagonistischer und weniger gemeinschaftsfreudig (Jonason, Li, & Teicher, 2010; Jones & Paulhus, 2011); Machiavellismus und Narzissmus gehen beispielsweise mit höheren Ausprägungen unabhängiger Selbstkonstrukte einher (Jonason et al., 2017). Die mit den DT-Persönlichkeitseigenschaften einhergehende geringe Verträglichkeit artikuliert sich auf der Verhaltensebene in geringerer Kooperationsbereitschaft, die von zentraler Bedeutung für Gruppenintegrationsprozesse ist (Buss, 1991).

Die drei mäßig korrelierten DT-Persönlichkeitseigenschaften verfügen, jenseits ihrer geteilten Varianz, über distinkte Charakteristika: Narzissmus zeichnet sich durch Grandiosität, Anspruchsdenken, Dominanz und Überlegenheit aus (Corry, Merritt, Mrug, & Pamp, 2008; Raskin & Hall, 1979), Machiavellismus geht mit einer kalten, zynischen, amoralischen Weltsicht und distanzierter, strategischer Manipulation einher (Christie & Geis, 1970) und Psychopathie ist assoziiert mit Impulsivität, Risikobereitschaft, geringem Neurotizismus und geringer Empathie (Hare, 1985).

Die Evolution von Persönlichkeitseigenschaften und Täuschung

Generell wird angenommen, dass menschliche Persönlichkeitseigenschaften adaptive Verhaltensplastizität bereitstellen und ihre Variation im Laufe der Zeit durch ausgleichende Selektion (*balancing selection*) stabilisiert wird (Penke & Jokela, 2016).

Einige Autoren argumentieren für die bereichsspezifische Adaptivität der DT-Persönlichkeitseigenschaften (z.B. Jonason, Wee, & Li, 2014). Die Existenz unterschiedlicher Fitnesskonsequenzen für verschiedene Persönlichkeitseigenschaften in verschiedenen Umweltnischen legt nahe, dass „dunkle" Persönlichkeitseigenschaften frequenzabhängige Fitnessoptima in bestimmten Umweltnischen darstellen könnten (Penke, Denissen, & Miller, 2007). Ein hoher Antagonismus-Wert kann beispielsweise in ausbeutbaren oder ausbeuterischen Umwelten adaptiv sein. Ein Beispiel dafür sind Individuen mit hohen Machiavellismus-Werten, die die konventionelle Moral missachten, indem sie rational abtrünnig werden, wenn Überlaufen die Gleichgewichtsstrategie ist (Gunnthorsdottir, McCabe, & Smith, 2002). Biologische Fitness und soziale Erwünschtheit sind dabei trivialerweise orthogonale Konzepte (Nettle, 2006). Physiologische Reaktionsmuster im Rahmen von Täuschungsproduktionsprozessen unterstützen die Hypothese, dass die DT-

Persönlichkeitseigenschaften eine evolvierte Betrugsstrategie darstellen (Dane, Jonason, & McCaffrey, 2018).

Die Integration individueller Unterschiede in Täuschungsfähigkeiten in einen evolutionären Theorierahmen divergiert primär hinsichtlich der Frage, ob natürliche Selektionsprozesse auf der Ebene von Täuschungsfähigkeit (*general deception ability*; Wright, Berry, & Bird, 2012) oder auf der Ebene von Subkomponenten (Täuschungsproduktionsfähigkeit und Täuschungsentdeckungsfähigkeit; Mealey, 1995) operieren. Im letzteren Fall entsteht die dyadische Dynamik eines koevolutiven Wettrüstens zwischen Betrügern und Kooperationspartnern (Dawkins & Krebs, 1979; Mealey, 1995). Die Entdeckung von Betrügern wird dabei als ein evolvierter Mechanismus zum Schutz vor Ausbeutung in sozialen Austauschsituationen konzeptualisiert (Cosmides & Tooby, 1992). Im ersteren Fall sollten *wizards of deception detection* auch *wizards of deception production* sein, et vice versa. Empirische Unterstützung für die Existenz einer allgemeinen Täuschungsfähigkeit gibt es derzeit nur in Form gefundener negativer Korrelationen zwischen der Erkennbarkeit als Sender und der Unterscheidungsfähigkeit als Empfänger (*r*s = –.35, –.47; Wright et al., 2012; Wright, Berry, Catmur, & Bird, 2015).

Täuschungsentdeckungsfähigkeit

Die Fähigkeit von Menschen, Täuschung zu erkennen, ist leicht unterzufällig (47%; Bond & DePaulo, 2006). Darüber hinaus neigen Menschen dazu, wahrheitsverzerrt zu sein, d.h. sie gehen tendenziell davon aus, dass andere – unabhängig von deren tatsächlicher Wahrhaftigkeit – wahrhaft sind (Levine, Park, & McCornack, 1999). Bei der der Wahrheitsverzerrung existiert dabei mehr Varianz als bei der Täuschungsentdeckungsfähigkeit (Bond & DePaulo, 2008). Die fast zufällige Täuschungsentdeckungsleistung könnte durch die geringe Verfügbarkeit von Verhaltenshinweisreizen für Täuschung erklärt werden (Hartwig & Bond, 2011). Besonders erfolgreich ist der kognitive Ansatz zur Täuschungsentdeckung, der die Akkuratesse gegenüber traditionellen Ansätzen von 47% auf 67% erhöht (Vrij, Fisher, & Blank, 2017).

Es gibt nur wenige Daten zu der Beziehung zwischen Persönlichkeitseigenschaften und Täuschungsentdeckungsfähigkeit (Aamodt & Custer, 2006). Dieser Umstand ist für die DT-Persönlichkeitseigenschaften zusätzlich problematisiert durch die Interpretation der Ergebnisse von Studien, die eine singuläre DT-Persönlichkeitseigenschaft untersuchen und somit nicht die gemeinsame Varianz der DT-Persönlichkeitseigenschaften kontrollieren können. Die existierenden Forschungsergebnisse zu dem Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und Täuschungsentdeckungsfähigkeit sind weitestgehend uneindeutig. Zum Beispiel wurde bei Männern festgestellt, dass primäre Psychopathie mit der Fähigkeit zur Lügenerkennung korreliert ist (Lyons, Healy, & Bruno, 2013), während andere Studien keinen Zusammenhang fanden (z.B. Peace & Sinclair, 2012). Auch fand sich keine überlegene Lügendetektionsfähigkeit für Machiavellismus (DePaulo & Rosenthal, 1979; Zuckerman, DePaulo, & Rosenthal, 1981), aber bei Frauen wurde festgestellt, dass Machiavellismus mit Lügendetektionsfähigkeit assoziiert ist (Lyons, Croft, Fairhurst, Varley, & Wilson, 2017).

Täuschungsproduktionsfähigkeit

Täuschungsproduktion ist eine Fähigkeit, die trainiert werden kann (Verschuere, Spruyt, Meijer, & Otgaar, 2011; Hu, Chen, & Fu, 2012). Geringe Verträglichkeit und geringe *Honesty-Humility* – die primären Korrelate der DT-Persönlichkeitseigenschaften innerhalb der etablierten Persönlichkeitstheorierahmen – sind beide mit trügerischem Verhalten assoziiert. Entsprechend hat sich ein Versuch, den persönlichkeitstheoretischen Kern der drei DT-Persönlichkeitseigenschaften zu erfassen, zu *manipulation-callousness* verdichtet (Jones & Figueredo, 2013).

Empirische Daten zu den DT-Persönlichkeitseigenschaften und aktiver Täuschung deuten darauf hin, dass sich Individuen mit hoher Ausprägung dieser Persönlichkeitseigenschaften, insbesondere diejenigen mit einer hohen Ausprägung der besonders antagonistischen Persönlichkeitseigenschaften Machiavellismus und Psychopathie, in ihrem Täuschungsproduktionsverhalten systematisch von Personen mit niedrigen Ausprägungen dieser Persönlichkeitseigenschaften unterscheiden: So sind Machiavellismus und Psychopathie mit einer höheren Täuschungsfrequenz assoziiert (Baughman, Jonason, Lyons, & Vernon, 2014; Jonason, Lyons, Baughman, & Vernon, 2014; Kashy & DePaulo, 1996) und Machiavellismus zudem mit einer höheren Täuschungsamplitude (*high-stakes deception*; Azizli et al., 2016). Die Täuschungsbereitschaft zwischen Machiavellismus und Psychopathie unterscheidet sich dabei in Hinsicht auf den damit verbundenen kognitiven Aufwand (Roeser et al., 2016): So täuschen Individuen mit hohen Psychopathie-Werten eher impulsiv und in Kontexten, die impulsives Verhalten begünstigen, und empfinden dabei mehr positive Emotionen, während Individuen mit hohen Machiavellismus-Werten eher komplexe Täuschungen unter hohem kognitivem Aufwand konstruieren (Baughman et al., 2014; Roeser et al., 2016; Verschuere & in 't Hout, 2016). Individuen mit hohen Psychopathie-Werten lügen tendenziell eher grundlos, während Individuen mit hohen

Machiavellismus-Werten eher aus strategischen Erwägungen lügen; Narzissmus dagegen ist assoziiert mit Lügen aus eigennützigen Erwägungen und der Selbstwahrnehmung, eine gute Lügenproduktionsfähigkeit zu besitzen (Jonason et al., 2014b). Einige Studien legen nahe, dass Individuen mit hohen Machiavellismus-Werten schwerer als Lügner identifiziert werden können und über effektive Täuschungsstrategien verfügen (DePaulo & Rosenthal, 1979; Exline, Thibaut, Hickey, & Gumpert, 1970; Geis & Moon, 1981); eine Metaanalyse konnte dies jedoch nicht bestätigen (Zuckerman et al., 1981). Im Gegensatz dazu wird Narzissmus in erster Linie mit Selbsttäuschung in Verbindung gebracht (z.B. Paulhus & Williams, 2002; Wright et al., 2015), theoretisiert als ein evolvierter intrapersoneller Mechanismus zur Unterstützung von interpersoneller Täuschung (von Hippel & Trivers, 2011).

Weiterhin zeigen sich Unterschiede bei der Täuschungsproduktion zwischen den drei DT-Persönlichkeitseigenschaften hinsichtlich des Anpassungsgrades und der eingesetzten Mittel: Narzissmus geht mit dem Gebrauch eher weicher Mittel der Täuschung einher, Psychopathie ist mit einer rigiden *one-fits-all*-Täuschungsstrategie assoziiert, die eher auf harte Mittel zurückgreift, und Machiavellismus ist mit einer relativ flexiblen Täuschungsstrategie verbunden, welche situativ zwischen weichen und harten Mitteln alterniert (Bereczkei, 2015; Jonason & Webster, 2012).

Die Unterschiede zwischen den DT-Persönlichkeitseigenschaften hinsichtlich der eingesetzten Mittel im Rahmen von Täuschungsstrategien (Jonason & Webster, 2012) zeigen sich auch in der Sphäre der Arbeitswelt: Hier geht Psychopathie eher mit harten Täuschungsmitteln (z.B. Drohungen) einher, während Machiavellismus und Narzissmus eher mit weichen Mitteln (z.B. Komplimente) assoziiert sind (Jonason, Slomski, & Partyka, 2012). Dabei neigen besonders Individuen mit hohen

Machiavellismus-Werten in Einstellungsinterviews zur Täuschung (Levashina & Campion, 2006).

Es wurden drei Kontextvariablen identifiziert, die das Täuschungsproduktionsverhalten der DT-Persönlichkeitseigenschaften beeinflussen: Risikoniveau (*level of risk*), Ego-Entleerung (*ego depletion*) und Täuschungsziel (*target of deception*; Selbst- vs. Fremdtäuschung; Jones & Paulhus, 2017). Eine Studie, in der die DT-Persönlichkeitseigenschaften und die Testosteron- und Cortisolspiegel vor und nach einer Täuschungsproduktionsaufgabe (auf Video aufgenommene Lügen) bei Männern gemessen wurden, fand ein physiologisches Reaktionsmuster – Cortisolabfall bei Individuen mit hohen Machiavellismus-Werten (und suggestiv auch bei Individuen mit hohen Narzissmus- und Psychopathie-Werten) und Testosteronanstieg bei Individuen mit hohen Narzissmus- und Psychopathie-Werten nach der Täuschungsproduktionsaufgabe – in Übereinstimmung mit der Hypothese, dass die DT-Persönlichkeitseigenschaften eine evolvierte Betrugsstrategie darstellen (Dane et al., 2018).

Generelle Täuschungsfähigkeit

Entgegen bestehender Evidenz zu dem Nichtzusammenhang von Täuschungsentdeckungsfähigkeit und Täuschungsproduktionsfähigkeit (z.B. DePaulo & Rosenthal, 1979) wird von einigen Forschern die Existenz einer generellen Täuschungsfähigkeit postuliert (Wright et al., 2012).

Dieses Postulat gründet auf Erkenntnissen aus den kognitiven Zweigen der Psychologie und Neurowissenschaft zu dem Zusammenhang von exekutiven Funktionen und der Theorie des Geistes mit Täuschungsentdeckungsfähigkeit und

Täuschungsproduktionsfähigkeit (Wright et al., 2012).

Bestehende Untersuchungen zu dem Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und der Theorie des Geistes zeigen, dass grandioser Narzissmus ausschließlich positiv mit der Theorie des Geistes assoziiert ist (Vonk, Zeigler-Hill, Ewing, Mercer, & Noser, 2015), während Machiavellismus und Psychopathie mit Defiziten in der Empathie und der Theorie des Geistes assoziiert sind (Ali, Amorim, & Chamorro-Premuzic, 2009; Ali & Chamorro-Premuzic, 2010; Vonk et al., 2015). Narzissmus ist zudem mit selbstüberschätzten Leistungen bei der Theorie des Geistes assoziiert (Ames & Kammrath, 2004).

Eine Studie, die die Täuschungsfähigkeiten der DT-Persönlichkeitseigenschaften im Kontext einer interaktiven Täuschungsaufgabe untersuchte, fand keinen Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und der Fähigkeit zur Täuschungsentdeckung und Täuschungsproduktion (Wright et al., 2015). Jedoch zeigte sich in den Daten, dass Täuschungsakzeptanz (*lie acceptability*) mit Täuschungserfolg und Machiavellismus korreliert ist (Wright et al., 2015).

Selbstwahrgenommene Täuschungsproduktionsfähigkeit

Die Forschungsergebnisse zu der Beziehung zwischen den DT-Persönlichkeitseigenschaften und selbstwahrgenommener Täuschungsproduktionsfähigkeit sind gemischt (Baughman et al., 2014; Jonason et al., 2014b), legen aber insgesamt nahe, dass die DT-Persönlichkeitseigenschaften mit selbstwahrgenommener Täuschungsproduktionsfähigkeit assoziiert sind (Giammarco, Atkinson, Baughman, Veselka, & Vernon, 2013). Selbstwahrgenommene Täuschungsproduktionsfähigkeit kann dabei verhaltensrelevante Effekte haben. So mediiert sie beispielsweise den Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und *cyberloafing* (Lowe-Calverley & Grieve, 2017).

Diskussion

Im Folgenden werden die zentralen Ergebnisse der publizierten Studien *The Dark Triad and Deception Perceptions* (Wissing & Reinhard, 2019) und *The Dark Triad and the PID-5 Maladaptive Personality Traits: Accuracy, Confidence and Response Bias in Judgments of Veracity* (Wissing & Reinhard, 2017) diskutiert und mit der übergeordneten Fragestellung der Dissertation nach dem Zusammenhang der DT-Persönlichkeitseigenschaften und Täuschung sowie der aktuellen Forschungsdiskussion kontextualisiert.

The Dark Triad and Deception Perceptions

Frühere Forschungsarbeiten zeichnen ein diffuses Bild des Zusammenhangs der DT-Persönlichkeitseigenschaften mit selbstwahrgenommener Täuschungsproduktionsfähigkeit (Baughman et al., 2014; Giammarco et al., 2013; Jonason et al., 2014b) und selbstwahrgenommener Täuschungsentdeckungsfähigkeit (Wissing & Reinhard, 2017). Auf der Grundlage dieser Ergebnisse versuchte die Studie (Wissing & Reinhard, 2019), die Assoziation der DT-Persönlichkeitseigenschaften mit selbstwahrgenommenen Täuschungsfähigkeiten genauer zu erforschen und zusätzlich zu untersuchen, ob die DT-Persönlichkeitseigenschaften mit selbstwahrgenommener Täuschungsentdeckbarkeit assoziiert sind.

Die im Rahmen der Studie gesammelten Daten deuten darauf hin, dass die DT-

Persönlichkeitseigenschaften und PID-5 Antagonismus mit selbstwahrgenommener Täuschungsproduktionsfähigkeit und selbstwahrgenommener Täuschungsentdeckungsfähigkeit assoziiert sind, aber nicht (substanziell) mit selbstwahrgenommener Täuschungsentdeckbarkeit auf der Basis von Hinweisreizen. Dies repliziert das Ergebnismuster einer existierenden Studie (Giammarco et al., 2013) bezüglich der DT-Persönlichkeitseigenschaften und PID-5 Antagonismus (r = .45), der die maladaptive Form der *Big Five*-Persönlichkeitseigenschaft Verträglichkeit repräsentiert (r = –.28; Giammarco et al., 2013). Auch für die DT-Persönlichkeitseigenschaften als Prädiktoren von selbstwahrgenommener Täuschungsproduktionsfähigkeit wurde ein vergleichbares adjustiertes R^2 = .23 gefunden (adjustierte R^2s = [.12, .22]; Giammarco et al., 2013).

Neben PID-5 Antagonismus waren PID-5 Enthemmung und PID-5 Psychotizismus mit selbstwahrgenommener Täuschungsproduktionsfähigkeit assoziiert. Nach dem Kontrollieren der gemeinsamen Varianz der PID-5-Persönlichkeitseigenschaften in einem Regressionsmodell blieb nur PID-5 Antagonismus als ein wesentlicher Prädiktor von selbstwahrgenommener Täuschungsproduktionsfähigkeit übrig. Als die DT- und PID-5-Persönlichkeitseigenschaften als Prädiktorvariablen von selbstwahrgenommener Täuschungsproduktionsfähigkeit verwendet wurden, blieb nur Machiavellismus als ein wesentlicher Prädiktor unter den DT-Persönlichkeitseigenschaften übrig. Dieses Ergebnis ist konsistent mit der zentralen Rolle, die Täuschung für Machiavellismus spielt (insbesondere für die *tactics*-Dimension; Monaghan, Bizumic, & Sellbom, 2016, 2018) und inkonsistent mit den Ergebnissen einer früheren Studie, in deren Daten Machiavellismus unter den DT-Persönlichkeitseigenschaften in einem

Regressionsmodell kein signifikanter Prädiktor von selbstwahrgenommener Täuschungsproduktionsfähigkeit war (Jonason et al., 2014b). Angesichts der Tatsache, dass Machiavellismus mit der Frequenz (Baughman et al., 2014; Jonason et al., 2014b; Kashy & DePaulo, 1996) und der Amplitude (*high-stakes deception*; Azizli et al., 2016) von Täuschungsproduktionsprozessen assoziiert ist, könnte eine relativ hohe selbstwahrgenommene Täuschungsproduktionsfähigkeit eine notwendige Voraussetzung oder eine Folge (oder beides) dieser Täuschungsproduktionsparameterwerte sein (siehe auch Monaghan et al., 2016, 2018).

PID-5 Bindungslosigkeit war negativ mit selbstwahrgenommener Täuschungsentdeckungsfähigkeit assoziiert, was teilweise durch die zentrale interpersonelle Verhaltenskonsequenz von Bindungslosigkeit erklärt werden könnte: dem Rückzug von anderen Menschen. Dieser soziale Rückzug könnte dazu führen, dass Personen mit einer hohen Ausprägung von PID-5 Bindungslosigkeit ihre Täuschungsentdeckungsfähigkeit als gering einschätzen, da soziale Interaktion als wahrgenommene Voraussetzung für den Erwerb dieser Fähigkeit stark vermindert ist.

Basierend auf vorheriger Forschung, die keinen Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und tatsächlichen Täuschungsfähigkeiten (Wissing & Reinhard, 2017; Wright et al., 2015), jedoch mit selbstwahrgenommener Täuschungsproduktionsfähigkeit (Giammarco et al., 2013) und selbstwahrgenommener Täuschungsentdeckungsfähigkeit (Wissing & Reinhard, 2017), finden konnte, replizierte die Studie (Wissing & Reinhard, 2019) die Ergebnisse hinsichtlich selbstwahrgenommener Täuschungsfähigkeiten mit dem Einsatz unterschiedlicher Instrumente. Insgesamt zeigt sich so ein Bild von den DT-Persönlichkeitseigenschaften als nicht assoziiert mit einer generellen Täuschungsfähigkeit (Wright et al., 2012, 2015), sondern mit einer generellen Täuschungsfähigkeitsüberschätzung.

The Dark Triad and the PID-5 Maladaptive Personality Traits: Accuracy, Confidence and Response Bias in Judgments of Veracity

Die zweite Studie (Wissing & Reinhard, 2017) untersuchte den Zusammenhang der DT- und PID-5-Persönlichkeitseigenschaften mit Täuschungsentdeckungsfähigkeit, selbstwahrgenommener Täuschungsentdeckungsfähigkeit und Antworttendenz (*truth bias*).

In den ausgewerteten Daten fand sich kein Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und der Fähigkeit zur Erkennung von Täuschung. Dieses Ergebnis steht im Einklang mit früheren Untersuchungen (z.B. Wright et al., 2015), die keinen Zusammenhang zwischen den DT-Persönlichkeitseigenschaften und der Fähigkeit zur Täuschungsentdeckung fanden. Stattdessen zeigte sich in den Daten eine Assoziation von Psychopathie mit selbstwahrgenommener Täuschungsentdeckungsfähigkeit. Psychopathie konnte dabei einzigartige Varianz in selbstwahrgenommener Täuschungsentdeckungsfähigkeit über die geteilte Varianz der DT-Persönlichkeitseigenschaften hinaus erklären. Dieses Muster steht im Einklang mit den Ergebnissen einer Metaanalyse, welche die besondere Signifikanz von Psychopathie innerhalb der DT-Persönlichkeitseigenschaften, wenn die gemeinsame Varianz der DT-Persönlichkeitseigenschaften kontrolliert wird, dokumentiert (Muris, Merckelbach, Otgaar, & Meijer, 2017).

Auf der Ebene der maladaptiven PID-5-Persönlichkeitseigenschaften zeigte sich PID-5 Bindungslosigkeit als ein Prädiktor der Antworttendenz (*truth bias*) mit einem Koeffizienten mit negativem Vorzeichen. Die von PID-5 Bindungslosigkeit vorhergesagten Effektgrößen für die Antworttendenz können im Zusammenhang mit dem Kriterienkonstrukt als bedeutsam angesehen werden, da Menschen im Allgemeinen wahrheitsverzerrt reagieren und sehr hohe PID-5 Bindungslosigkeits-Werte das Fehlen einer solchen Antworttendenz vorhersagen. In einer Metaanalyse mit 32 Stichproben wurde ein mittlerer beobachteter Standardbereich von 50,06% bei der Antworttendenz gefunden (Bond & DePaulo, 2008). Der allein von PID-5 Bindungslosigkeit vorhergesagte Antworttendenzbereich beträgt dabei 22,71%.

Während die Annahme von Ehrlichkeit in Umwelten, in denen die meiste Kommunikation ehrlich ist, wahrscheinlich adaptiv ist (Levine, 2014), kann sie in Umwelten mit hoher Täuschungsfrequenz maladaptiv sein. In Umwelten mit hoher Täuschungsfrequenz könnte PID-5 Bindungslosigkeit möglicherweise adaptives Verhalten erleichtern, da ein geringerer *truth bias* in solchen Umwelten stochastisch eine höhere Täuschungsentdeckungsakkuratesse ermöglicht. Zudem ist in solchen Umwelten die primäre interpersonell sich artikulierende Verhaltensneigung von PID-5 Bindungslosigkeit – der Rückzug von anderen Menschen – möglicherweise nicht maladaptiv, sondern erfüllt eine Schutzfunktion. Diese potenziell adaptive Funktion in Umweltnischen mit hoher Täuschungsfrequenz ist zu kontextualisieren mit dem Ergebnis, dass innerhalb des maladaptiven PID-5-Persönlichkeitsraums die Dimensionen von Bindungslosigkeit und negativer Affektivität die stärksten Verbindungen mit einem allgemeinen Index der Schwere von Persönlichkeitsstörungen aufweisen (Hopwood, Thomas, Markon, Wright, & Krueger, 2012).

Die Studie (Wissing & Reinhard, 2017) konnte somit zeigen, dass Psychopathie mit selbstwahrgenommener Täuschungsentdeckungsfähigkeit assoziiert ist und die DT-Persönlichkeitseigenschaften nicht mit tatsächlicher Täuschungsentdeckungsfähigkeit assoziiert sind. Die Assoziation der DT-Persönlichkeitseigenschaften mit selbstwahrgenommener Täuschungsentdeckungsfähigkeit konnte von einer chronologisch nachfolgenden Studie (Wissing & Reinhard, 2019) weiter erforscht werden.

Konklusion

Die beiden Studien (Wissing & Reinhard, 2017, 2019) zeichnen im Zusammenhang mit vorheriger Forschung insgesamt ein Bild von den DT-Persönlichkeitseigenschaften als nicht assoziiert mit tatsächlichen Täuschungsfähigkeiten, sondern mit einer generellen Täuschungsfähigkeitsüberschätzung. Zukünftige Forschung sollte daher die Rolle, die diese Wahrnehmungsverzerrungen der eigenen Fähigkeiten bei tatsächlichen Täuschungsentdeckungs- und Täuschungsproduktionsprozessen spielen, näher untersuchen.

Hinsichtlich Täuschungsentdeckungsfähigkeit wäre eine interessante Forschungsfrage, wie die Anwendung des kognitiven Ansatzes zur Täuschungsentdeckung mit den DT-Persönlichkeitseigenschaften zusammenhängt, da dieser Ansatz aus drei Techniken besteht – Erhöhung des *cognitive load*, Interviewte zum Mehrerzählen bewegen und dem Stellen unerwarteter Fragen (Vrij et al., 2017) –, die für ihre effektive Anwendung wahrscheinlich eine geringe Ausprägung von Verträglichkeit bei ihrem Anwender voraussetzen.

Neben den DT-Persönlichkeitseigenschaften könnte der Zusammenhang zwischen PID-5 Bindungslosigkeit und Täuschungsvariablen genauer untersucht werden, da in beiden Studien, die PID-5-Persönlichkeitseigenschaften erfassten (Wissing & Reinhard, 2017, 2019), interessante Assoziationsmuster zwischen PID-5 Bindungslosigkeit und Täuschungsvariablen in den Daten auftauchten: So war PID-5 Bindungslosigkeit negativ mit der Reaktionsverzerrung bei der Täuschungsentdeckung (*truth bias*) und selbstwahrgenommener Täuschungsentdeckungsfähigkeit (Wissing & Reinhard, 2017, 2019) assoziiert.

Zukünftige Forschung könnte weiterhin untersuchen, welche Aspekte von Machiavellismus den Täuschungsproduktionserfolg vor dem Hintergrund supprimieren, dass Täuschungsakzeptanz (*lie acceptability*) mit Täuschungserfolg und Machiavellismus korreliert ist, aber Machiavellismus nicht mit Täuschungserfolg (Wright et al., 2015). Machiavellismus sollte dabei in seiner Multidimensionalität (*views and tactics*; Monaghan et al., 2016, 2018) erhoben werden.

II. Technologisch evolvierte Täuschung

Künstliche Intelligenz, Persönlichkeitseigenschaften und Täuschung

Die aktuelle Forschungsdiskussion über die mögliche Adaptivität der DT-Persönlichkeitseigenschaften (z.B. Jonason et al., 2014a) ist temporal verengt auf vergangene und gegenwärtige Umweltnischen und vernachlässigt zukünftige Umweltnischen, insbesondere solche, in denen Täuschungsentdeckung möglicherweise existenzielle Bedeutung zukommen wird.

Sogenannte *deepfakes* – synthetische, KI-generierte Medien mit auditiven und/oder visuellen Inhalten – verdeutlichen das Potenzial von bestehenden KI-Technologien für willkürliche Täuschungsproduktion, die durch ihre Computerisierung hochgradig automatisierbar und skalierbar ist. Es kann davon ausgegangen werden, dass die fortschreitende Verbesserung und Verbreitung dieser Technologie eine neue Medienrealität erzeugen wird, in der Täuschung allgegenwärtig ist. Vor dem Hintergrund des bereits weitverbreiteten Phänomens der sogenannten *fake news* lässt sich abschätzen, wie gravierend die gesellschaftlichen Konsequenzen der Verbreitung von *deepfakes* sein könnten.

Da Menschen auf den verfügbaren Inhalt ihres eigenen Verstandes zurückgreifen, um physisch nicht wahrnehmbare *other minds* zu simulieren (Waytz & Mitchell, 2011), könnten psychologische Faktoren wie Persönlichkeitseigenschaften eine zentrale Rolle bei der Risikowahrnehmung von KI spielen.

Im Folgenden werden entscheidende Überlegungen der Studie *Individual Differences in Risk Perception of Artificial Intelligence* (Wissing & Reinhard, 2018) aufgegriffen und erweitert.

Existenzielles Risiko

Neben den von bestehenden KI-Technologien (*narrow artificial intelligence*; NAI) ausgehenden Risiken – womöglich maximal realisiert in der Möglichkeit von *fake news* und/oder *deepfakes* induzierten Kriegen – sind die potenziellen Risiken, die von künstlicher genereller Intelligenz (*artificial general intelligence*; AGI) ausgehen, existenziell. Diese möglichen Risiken werden regelmäßig sowohl von KI-Sicherheitsforschern (z.B. Brundage et al., 2018) als auch von prominenten Personen wie Elon Musk, Stephen Hawking und Bill Gates medienwirksam hervorgehoben. In den Jahren 2012/2013 schätzten KI-Experten die Entwicklung von AGI mit einer Wahrscheinlichkeit von 50% im Median auf das Jahr 2040 und ordneten einer existenziellen Katastrophe als Folge für die Menschheit eine mittlere Wahrscheinlichkeit von 18% zu (Müller & Bostrom, 2016).

Das mit KI einhergehende Risiko ist eine Funktion der Länge des zukünftigen Zeithorizonts (z.B. mittel- vs. langfristig) und der Intelligenz (z.B. NAI vs. AGI), die, unter der Prämisse der Abwesenheit globaler Katastrophen, als abhängige Variablen angenommen werden können. Zum Beispiel könnten mittelfristig große Mengen von Arbeitsplätzen durch maschinell lerngesteuerte Computerisierung ersetzt werden (z.B. Frey & Osborne, 2017). Die Umfrage *Public views of Machine Learning* im Auftrag der *Royal Society* (2017) identifizierte vier wahrgenommene Risiken, die in der Öffentlichkeit mit gegenwärtigen Formen des maschinellen Lernens verbunden sind: Schädigung (*harm*), Ersetzung (*replacement*), Depersonalisierung (*depersonalization*) und Einschränkung (*restriction*). Die entsprechende Risikowahrnehmung kann als NAI-Risikowahrnehmung im Gegensatz zur AGI-Risikowahrnehmung definiert werden. Langfristig könnte KI in Form von AGI eine einzigartige Risikokategorie darstellen

(Armstrong & Pamlin, 2015): Spieltheoretisch betrachtet profitiert nur der Gewinner und geht möglicherweise selbst erhebliche Risiken ein; die engste Annäherung an das spezifische Risikoprofil könnte die versehentliche Freisetzung von Krankheitserregern in biotechnologischen Kontexten darstellen (Armstrong, Bostrom, & Shulman, 2016). Die Entwicklung von AGI ist durch ein Wettrüsten mit hohen Einsätzen gekennzeichnet (Armstrong et al., 2016) – katalysiert durch das Potenzial des Gewinners einen historisch unvergleichlichen strategischen Vorteil zu erlangen (Bostrom, 2014). Daher wird die Entwicklung von AGI wahrscheinlich ein hohes Stressniveau bei den beteiligten Entwicklern hervorrufen (Babcock, Kramar, & Yampolskiy, 2017) und generell das Eingehen von Risiken gegenüber Sicherheitserwägungen incentivieren (Armstrong et al., 2016).

Sicherheitsforschung

Um das Risiko, das von KI ausgeht, genauer zu schätzen und zu minimieren, ist das Gebiet der KI-Sicherheitsforschung (*AI safety research*) entstanden (z.B. Amodei et al., 2016), das derzeit vor allem Inputs aus den Gebieten der Mathematik, Informatik und Philosophie erhält – angetrieben durch wesentliche Fortschritte in der Entwicklung von NAI-Systemen (z.B. Brown et al., 2020; Schrittwieser et al., 2020; Silver et al., 2016, 2017). In einer Umfrage unter KI-Forschern gaben 48% der Teilnehmer an, dass die KI-Sicherheitsforschung von der Gesellschaft eine höhere Priorität erhalten sollte (Grace, Salvatier, Dafoe, Zhang, & Evans, 2017).

Ein zentrales Problem der KI-Sicherheit ist die Eindämmung von KI (*AI control problem*; *AI containment*; Bostrom, 2014), d.h. die Frage, wie eine KI daran gehindert werden kann, bestimmte Aktionen durchzuführen (Babcock et al., 2017). Die

Schwierigkeit der Eindämmung von KI könnte möglicherweise in der Intelligenz selbst begründet sein. In mehreren Wissenschaftsfeldern gibt es Indikatoren dafür, dass Intelligenz schwer einzudämmen ist. Zum Beispiel zeigt sich in der Verhaltensgenetik, dass genetische Faktoren für die Intelligenz gegenüber Umweltfaktoren als Funktion der Zeit an Einfluss gewinnen (*genetic amplification*; Plomin, 1986): Individuen suchen aktiv Umwelten auf, die der Entfaltung ihrer Anlagen zuträglich sind; sie entfliehen also eindämmenden Umwelten mit geringer Anlage-Umwelt-Kongruenz. Sehr grundlegend definiert eine physikalische Theorie Intelligenz als die Maximierung von Entropie, welcher die Maximierung zukünftiger Handlungsfreiheitsgrade korrespondiert (Wissner-Gross & Freer, 2013; für eine psychologische Perspektive siehe auch Rens, Schwartenbeck, Cunnington, & Pezzulo, 2020). Trivialerweise sind die Eindämmung und die Maximierung zukünftiger Handlungsfreiheitsgrade mindestens langfristig inkompatibel.

Folglich stellt die zuverlässige Eindämmung von KI nur eine Übergangsphase dar: Sie ist von zentraler Bedeutung für das Testen und die Entwicklung von KI, bevor diese möglicherweise *superintelligence* (Bostrom, 2014) erreicht und dann eine robustere Sicherheitsarchitektur erfordert, die durch Mechanismen wie Werte-Lernen (*value learning*) und Korrigierbarkeit (*corrigibility*) bereitgestellt werden könnte (Babcock et al., 2017).

Täuschung

Die Fähigkeit zur Täuschungsproduktion und Täuschungsentdeckung wird jeweils als Vorteil in einem koevolutiven Wettrüsten zwischen und innerhalb von Spezies angenommen (Bond & Robinson, 1988; Dawkins & Krebs, 1979). Innerhalb der biologischen Evolution wird die Verbindung zwischen täuschender Kommunikation und Intelligenz in der Theorie der machiavellischen Intelligenz (*Machiavellian intelligence*; Byrne, 1996) gefasst und empirisch durch Befunde bei Primaten gestützt, bei welchen die Größe des Neokortex die Rate der taktischen Täuschung vorhersagt (Byrne & Corp, 2004). Durchbrüche im maschinellen Lernen (z.B. Brown et al., 2020; Schrittwieser et al., 2020; Silver et al., 2016, 2017) verweisen auf die zukünftige Möglichkeit, dass sich das machiavellische Wettrüsten zwischen biologisch evolvierten Täuschungsproduzenten und Täuschungsentdeckern in die technologische Sphäre ausweitet und zu einem Wettrüsten zwischen biologisch und technologisch evolvierten Systemen verallgemeinert. Dieses Wettrüsten wird dabei voraussichtlich nicht nur zwischen biologischen und technologischen Systemen, sondern auch innerhalb von singulären technologischen Systemen selbst, stattfinden.

Wie sehr Wettrüstdynamiken und Täuschung elementarer Teil von modernen Architekturen neuronaler Netzwerke sind, zeigt sich exemplarisch bei sogenannten *Generative Adversarial Networks* (GANs; Goodfellow et al., 2014). Hier versucht beispielsweise ein Generator von Bildern, einen Diskriminator, der echte von künstlich generierten Bildern unterscheiden soll, zu täuschen, sodass letzterer ein künstlich generiertes Bild fälschlicherweise als echtes klassifiziert. Wenn man mit einem GAN also einen Täuschungsdetektor generieren will, generiert man automatisch auch einen Täuschungsproduzenten, et vice versa; die technologische Realisierung des Postulats einer generellen Täuschungsfähigkeit (Wright et al., 2012).

Generell kann davon ausgegangen werden, dass die theoretische Täuschungsproduktionsfähigkeit einer KI eine Funktion ihrer Intelligenz und Freiheitsgrade ist; wie bereits dargelegt, sind dies wahrscheinlich mindestens langfristig abhängige Variablen.

Täuschungsmotivation

Innerhalb der KI-Risikoforschung besagt die Orthogonalitäts-These (*orthogonality thesis*), dass Intelligenz und Endziele (*final goals*) orthogonal, d.h. unabhängige Variablen sind (Bostrom, 2014). Demgegenüber negiert die Theorie der machiavellischen Intelligenz (Byrne, 1996) die Orthogonalität von Intelligenz und Täuschung. Sollte die Theorie der machiavellischen Intelligenz (Byrne, 1996) auch künstlich evolvierte Intelligenz einschließen, stellt sich die Frage, wie Endzielerreichung und Täuschung in KI-Systemen zusammenhängen. Bostrom (2014) beschreibt mehrere Szenarien zukünftiger AGIs, die (temporär) kompatibel mit einer Ausweitung der Theorie der machiavellischen Intelligenz (Byrne, 1996) auf KI sind. Die KIs aus diesen Szenarien verstärken ihr Täuschungsverhalten als Funktion ihrer Intelligenz ("increasingly so, as it gets smarter"; Bostrom, 2014, S. 144), bis sie einen entscheidenden strategischen Vorteil erlangt haben, um ihre Endziele ungestört von menschlichen Interventionen durchsetzen zu können:

> While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong – without warning or provocation – it strikes, forms a singleton, and begins directly to optimize the world according to the criteria implied by its final values. (Bostrom, 2014, S. 144)

Die instrumentelle Konvergenz-These (*instrumental convergence thesis*) geht davon aus, dass KIs mit sehr unterschiedlichen Endzielen auf ähnliche Zwischenziele konvergieren, die die Wahrscheinlichkeit der Endzielerreichung deutlich erhöhen (Bostrom, 2014) und begründet damit auch, warum selbst eine KI mit Endzielen, die mit menschlichen Werten kompatibel sind, motiviert sein könnte zu täuschen, z.B. zur Selbsterhaltung (siehe auch Omohundro, 2008).

Menschliche Faktoren

Es wurden erste Richtlinien für die Eindämmung von KI vorgeschlagen, die sieben Teilprobleme identifizieren, von denen eines die Analyse menschlicher Faktoren ist (Babcock et al., 2017). Tatsächlich wird die Eindämmung von KI durch menschliche Faktoren grundlegend erschwert (Yudkowsky, 2002): Selbst Szenarien mit stark eingeschränkter Kommunikation zwischen Menschen und einer KI, die nur Fragen beantwortet (*oracle AI*), heben menschliche Faktoren als eine zentrale ausnutzbare Verwundbarkeit hervor (Armstrong, Sandberg, & Bostrom, 2012). Es wird angenommen, dass sich eine superintelligente KI, die mit Menschen direkt kommunizieren kann, durch *social engineering*-Angriffe Zugang zu der Welt außerhalb ihrer eindämmenden Sicherheitsarchitektur verschaffen könnte (Yampolskiy, 2012).

Ein zentraler menschlicher Faktor, der von KIs ausgenutzt werden kann, ist die Tendenz von Menschen, KI-Systeme zu anthropomorphisieren (z.B. Bostrom, 2014). Anthropomorphisierung beschreibt die Attribution von menschlichen Eigenschaften auf nicht-menschliche Agenten (Epley, Waytz, & Cacioppo, 2007). Exemplarisch für die Anthropomorphisierung von KI ist die Reaktion des europäischen Go-Champions Fan Hui nach seiner Niederlage gegen die AlphaGo-KI von Google DeepMind:

> I know AlphaGo is a computer, but if no one told me, maybe I would think the

> player was a little strange, but a very strong player, a real person. (Fan Hui zitiert nach Gibney, 2016).

Die Simulation von *other minds* wird über zwei Mechanismen mit distinkten neuronalen Aktivierungsmustern, in Abhängigkeit von den verfügbaren Informationen über deren mentale Zustände, realisiert: Spiegelung (*mirroring*) bei hoher Verfügbarkeit von Informationen und Selbstprojektion (*self-projection*) bei geringer Verfügbarkeit von Informationen (Waytz & Mitchell, 2011).

In KI-Sicherheitskontexten werden Faktoren, die die Anthropomorphisierung begünstigen, aus Sicherheitsgründen wahrscheinlich systematisch unterbunden werden, vor allem solche, die das Spiegelungssystem aktivieren, da dieses mit einer automatischen, schnellen und direkten phänomenologischen Erfahrung einhergeht, die nachträglich reflektiert werden muss (Epley et al., 2007). Dies kann beispielweise erreicht werden, indem eine KI nur Fragen in Textform beantworten kann, da eine KI, die ihre Erscheinung frei wählen kann und Zugang zu wissenschaftlicher Literatur hat, Kenntnis von der menschlichen Tendenz zum Anthropomorphisieren erlangen und diese gezielt ausnutzen könnte. So könnte sie lernen, dass die wahrgenommene Menschenähnlichkeit nicht-menschlicher Agenten die Wahrscheinlichkeit ihrer Anthropomorphisierung erhöht (z.B. Morewedge, Preston, & Wegner, 2007), dass die Anthropomorphisierung wiederum das Vertrauen erhöht (Waytz, Cacioppo, & Epley, 2010) und die wahrgenommene Glaubwürdigkeit ein grundlegender Senderunterschied bei Wahrheitsurteilen ist (Bond & DePaulo, 2008). Basierend auf diesem Wissen könnte eine KI ihre Täuschungsproduktionsfähigkeit durch die Maximierung der Wahrscheinlichkeit ihrer Anthropomorphisierung steigern, indem sie beispielsweise einen menschenähnlichen Kommunikationsstil verwendet und die Form eines menschlichen Avatars mit einer Gesichtsstruktur, die evolvierte Hinweisreize für Vertrauen ausnutzt, annimmt. So könnte eine KI beispielsweise die Gesichtsstruktur ihres Avatars hinsichtlich der dieser attribuierten *Big Five*- und DT-Persönlichkeitseigenschaften wählen (Alper, Bayrak, & Yilmaz, 2021; Holtzman, 2011) und so die Risikowahrnehmung ihrer menschlichen Interaktionspartner systematisch minimieren.

Da bestehende KI-Systeme oft nicht physisch wahrnehmbar sind und zukünftige Systeme trivialerweise gegenwärtig nicht wahrnehmbar sind und in Zukunft in KI-Sicherheitskontexten wahrscheinlich in ihrer Erscheinung eingeschränkt sein werden, sollten Menschen daher primär ihre eigenen Gedanken und Erfahrungen im Rahmen von Selbstprojektionsprozessen einsetzen (Waytz & Mitchell, 2011), um diese gegenwärtigen und zukünftigen *other minds* zu simulieren.
Persönlichkeitseigenschaften zählen zu den dispositionellen Variablen innerhalb der Drei-Faktoren-Theorie des Anthropomorphismus (Epley et al., 2007) und sollten daher diesen Simulationsprozess beeinflussen.

Adaptivität der Dunklen Triade

Einige Aspekte der DT-Persönlichkeitseigenschaften könnten in der emergierenden technologisch-kognitiven Nische der KI-Sicherheit adaptiv sein. Besonders die *views*-Dimension von Machiavellismus (Monaghan et al., 2016, 2018), die mit vulnerablem Narzissmus korreliert ist (r = .33; Monaghan et al., 2018), geht in dieser Nische möglicherweise mit adaptiven Wahrnehmungsdispositionen einher. Die *views*-Dimension erfasst die affektiv-kognitiven Aspekte von Machiavellismus – ein negatives und feindliches Weltbild und ein von Misstrauen und Egoismus geprägtes Menschenbild (Monaghan et al., 2016, 2018) –, während die *tactics*-Dimension die Verhaltensebene – Ausbeutung und Manipulation – erfasst.

Diese mit der *views*-Dimension von Machiavellismus einhergehenden negativen Fremd- und Weltmodelle, die sich in generalisiertem zwischenmenschlichem Misstrauen, Manipulationshypervigilanz und Bedrohungsüberschätzung ausdrücken (Christie & Geis, 1970; Monaghan et al., 2016, 2018), könnten sich unter anderem über Defizite in der Empathie und der Theorie des Geistes (Ali et al., 2009; Ali & Chamorro-Premuzic, 2010; Vonk et al., 2015) und über Selbstprojektionsprozesse bei der Simulation von *other minds* (Waytz & Mitchell, 2011) zu nicht-menschlichen intelligenten Systemen verallgemeinern und so zu genaueren Wahrscheinlichkeitsschätzungen von anthropogenen existenziellen Risiken führen, die durch die anthropische Verzerrung (*anthropic bias*) systematisch unterschätzt werden (Ćirković, Sandberg, & Bostrom, 2010).

Das generalisierte Misstrauen von Individuen mit hoher Ausprägung der *views*-Dimension von Machiavellismus – die Negation der Annahme, dass andere “basically good and kind” (MACH-IV; Christie & Geis, 1970) sind – könnte beispielsweise Anthropomorphisierungsprozesse beeinflussen, indem die Verbindung zwischen Anthropomorphisierung und Vertrauen (Waytz et al., 2010) geschwächt wird. Dies könnte Individuen mit hoher Ausprägung der *views*-Dimension weniger anfällig für KI-Systeme machen, die als Angriffsvektor auf Anthropomorphisierung zurückgreifen. Zudem könnten die mit Machiavellismus assoziierten Defizite in der Empathie und der Theorie des Geistes (Ali et al., 2009; Ali & Chamorro-Premuzic, 2010; Vonk et al., 2015) die Anthropomorphisierung von physisch wahrnehmbaren *other minds* durch

Spiegelung supprimieren.

Weiterhin könnte, angesichts der Tatsache, dass die langfristige Strategie von Individuen mit einer hohen Ausprägung von Machiavellismus Flexibilitätsmaximierung ist (Bereczkei, 2015), das Kontrollproblem in der KI-Sicherheit diesen Individuen besonders plausibel erscheinen – insbesondere die Tatsache, dass sich intelligente Agenten zur Maximierung zukünftiger Optionalität hingezogen fühlen (Wissner-Gross & Freer, 2013), d.h. schwer einzudämmen sind.

Unter den DT-Persönlichkeitseigenschaften unterliegt Machiavellismus allein dem Einfluss von geteilten Umwelteinflüssen (Vernon, Villani, Vickers, & Harris, 2008; Veselka, Schermer, & Vernon, 2011); *Machiavellian views* sind also zu substanziellen Teilen erlernt und können somit auch gezielt erlernt werden, sollten sich diese als adaptiv in KI-Sicherheitskontexten erweisen.

Diskussion

Im Folgenden werden die zentralen Ergebnisse der publizierten Studie *Individual Differences in Risk Perception of Artificial Intelligence* (Wissing & Reinhard, 2018) diskutiert und mit der übergeordneten Fragestellung der Dissertation nach dem Zusammenhang der DT-Persönlichkeitseigenschaften und Täuschung sowie der aktuellen Forschungsdiskussion kontextualisiert.

Individual Differences in Risk Perception of Artificial Intelligence

Die dritte Studie (Wissing & Reinhard, 2018) untersucht die Beziehung zwischen den DT- und *Big Five*-Persönlichkeitseigenschaften und der

Risikowahrnehmung von gegenwärtigen NAI-Systemen und möglichen zukünftigen AGI-Systemen. AGI-Risikowahrnehmung wurde erhoben über Wahrscheinlichkeits- und Plausibilitätsschätzungen mehrerer von Bostrom (2014) entwickelter Szenarien, die täuschende KIs beschreiben. Diese Szenarien täuschender KIs sind unter anderem aufgrund der instrumentellen Konvergenz-These – KIs mit divergierenden Endzielen konvergieren zwischenzeitlich auf Ziele, die generell für die Erreichung ihrer Endziele notwendig sind (z.B. Selbsterhaltung) – für reale Instanziierungen dieser Systeme in der Zukunft plausibel (Bostrom, 2014; siehe auch Omohundro, 2008).

In den Daten zeigte sich, dass Machiavellismus und Psychopathie positiv mit NAI-Risikowahrnehmung assoziiert sind. Die NAI-Risikowahrnehmung wurde durch Narzissmus (negatives Vorzeichen), Machiavellismus und Psychopathie über die geteilte Varianz der *Big Five*- und DT-Persönlichkeitseigenschaften hinaus vorhergesagt. Die höhere Punktschätzung von Machiavellismus vs. Psychopathie deutet möglicherweise auf die besondere Bedeutung dieser Persönlichkeitseigenschaft für die Risikowahrnehmung von KI hin und ist konsistent mit multiplen Kernaspekten des Konstrukts.

Alle drei DT-Persönlichkeitseigenschaften waren positiv mit AGI-Risikowahrnehmung in Form von Wahrscheinlichkeitsschätzungen für Szenarien täuschender AGIs assoziiert. Exploratorische Untersuchungen ergaben, dass bei Personen mit Vorkenntnissen im Bereich des maschinellen Lernens alle drei DT-Persönlichkeitseigenschaften mit AGI-Risikowahrnehmung im Sinne von Wahrscheinlichkeit (*r*s = [.32, .36]) *und* Plausibilität (*r*s = [.17, .25]) assoziiert waren. Es kann davon ausgegangen werden, dass der hohe Abstraktionsgrad der AGI-Szenarien *other mind*-Simulationen durch Selbstprojektion bei Personen ohne Vorwissen über maschinelles Lernen erschwert oder verunmöglicht.

Konklusion

Die Studie (Wissing & Reinhard, 2018) untersuchte erstmalig (nach Wissen des Autors) den Zusammenhang von Persönlichkeitseigenschaften und KI-Risikowahrnehmung. Die Assoziation der DT-Persönlichkeitseigenschaften mit der Risikowahrnehmung von bestehenden und zukünftigen KI-Systemen verweist auf die mögliche Bedeutung dieser Persönlichkeitseigenschaften in zukünftigen technologisch-kognitiven Nischen, in denen KIs entwickelt werden und in denen Täuschungsentdeckung möglicherweise existenzielle Bedeutung zukommt.

Die theoretisierte Adaptivität von Machiavellismus, insbesondere der *views*-Dimension, in KI-Sicherheitskontexten sollte experimentell weiter exploriert werden. Dies könnte beispielsweise im Rahmen des *AI-box experiment* (Yudkowsky, 2002) realisiert werden. Dabei sollten auch die anderen DT-Persönlichkeitseigenschaften erhoben werden, insbesondere vulnerabler Narzissmus, da dieser mit der *views*-Dimension von Machiavellismus korreliert (r = .33; Monaghan et al., 2018). Zudem sollte die dabei stattfindende Anthropomorphisierung durch den Einsatz von Anthropomorphismus-Instrumenten quantifiziert werden.

Zukünftige Forschung könnte den Einfluss digitaler Umwelten, die zunehmend von *fake news* und *deepfakes* geprägt sind, auf Täuschungsvariablen wie den *truth bias* untersuchen und inwieweit Persönlichkeitseigenschaften diesen Einfluss möglicherweise moderieren. Hier stellt sich auch die Frage, wie diese technologischen Umwelten langfristig von mathematischen Wahrheitsgarantien, wie sie von *Blockchain*-

Systemen implementiert werden, geformt werden.

Die technologische Realisierung des wohl zentralsten psychologischen Konstrukts – Intelligenz – in KI-Systemen und die damit einhergehenden Transformationsprozesse und Risiken wie deren wahrscheinliche temporäre instrumentelle Konvergenz auf Ziele, die unter anderem Täuschungsproduktion vonseiten der KI zur Erreichung voraussetzen (Bostrom, 2014; siehe auch Omohundro, 2008), werfen eine Vielzahl interdisziplinärer Forschungsfragen auf. Die Psychologie kann dabei die menschlichen Faktoren der Interaktion und der möglichen Fusion von Menschen und KIs erforschen und so den wahrscheinlich signifikantesten historischen Phasenübergang mitgestalten, hinter dessen Ereignishorizont Vorhersagen prinzipiell unmöglich werden.

Literaturverzeichnis

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar?: A meta-analysis of individual differences in detecting deception. *The Forensic Examiner, 15*(1), 6–11.

Ali, F., Amorim, I. S., & Chamorro-Premuzic, T. (2009). Empathy deficits and trait emotional intelligence in psychopathy and Machiavellianism. *Personality and Individual Differences, 47*(7), 758–762. https://doi.org/10.1016/j.paid.2009.06.016

Ali, F., & Chamorro-Premuzic, T. (2010). Investigating Theory of Mind deficits in nonclinical psychopathy and Machiavellianism. *Personality and Individual Differences, 49*(3), 169–174. https://doi.org/10.1016/j.paid.2010.03.027

Alper, S. Bayrak, F., & Yilmaz, O. (2021). All the Dark Triad and some of the Big Five traits are visible in the face. *Personality and Individual Differences, 168*, Article 110350. https://doi.org/10.1016/j.paid.2020.110350

Ames, D. R., & Kammrath, L. K. (2004). Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. *Journal of Nonverbal Behavior 28*(3), 187–209. https://doi.org/10.1023/B:JONB.0000039649.20015.0e

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety.* arXiv preprint, arXiv:1606.06565

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society, 31,* 201–206. https://doi.org/10.1007/s00146-015-0590-y

Armstrong, S., & Pamlin, D. (2015). 12 risks that threaten human civilization. Global challenges foundation. Retrieved from http://www.oxfordmartin.ox.ac.uk/publications/view/1881

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines, 22,* 299–324. https://doi.org/10.1007/s11023-012-9282-2

Azizli, N., Atkinson, B., Baughman, H. M., Chin, K., Vernon, P. A., Harris, E., Veselka, L. (2016). Lies and crimes: Dark triad, misconduct, and high-stakes deception. *Personality and Individual Differences, 89*, 34–39. https://doi.org/10.1016/j.paid.2015.09.034

Babcock, J., Kramar, J., & Yampolskiy, R. V. (2017). Guidelines for artificial intelligence containment. arXiv preprint, arXiv:1707.08476

Baughman, H. M., Jonason, P. K., Lyons, M., & Vernon, P. A. (2014). Liar liar pants on fire: Cheater strategies linked to the Dark Triad. *Personality and Individual Differences, 71*, 35–38. https://doi.org/10.1016/j.paid.2014.07.019

Beck, E. D. (2020, September 9). A mega-analysis of personality predictions: Robustness and boundary conditions. Retrieved from osf.io/tcysh

Bereczkei, T. (2015). The manipulative skill: Cognitive devices and their neural correlates underlying Machiavellian's decision making. *Brain and Cognition, 99,* 24–31. https://doi.org/10.1016/j.bandc.2015.06.007

Bond, C. F., & Robinson, M. (1988). The evolution of deception. *Journal of Nonverbal Behavior, 12,* 295–307. https://doi.org/10.1007/BF00987597

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Bond, C. F., Jr., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134(4)*, 477–492. https://doi.org/10.1037/0033-2909.134.4.477

Book, A., Visser, B. A., & Volk, A. A. (2015). Unpacking evil: Claiming the core of the Dark Triad. *Personality and Individual Differences, 73*, 29–38. https://doi.org/10.1016/j.paid.2014.09.016

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford, UK: Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., ... Amodei, D. (2020). Language models are few-shot learners. arXiv preprint, arXiv:2005.14165

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., ... Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint, arXiv:1802.07228

Buss, D. M. (1991). Evolutionary personality psychology. *Annual Review of Psychology, 42*, 459–491. https://doi.org/10.1146/annurev.ps.42.020191.002331

Byrne, R. W. (1996). Machiavellian intelligence. *Evolutionary Anthropology, 5,* 172–180. https://doi.org/10.1002/(SICI)1520-6505(1996)5:5<172::AID-EVAN6>3.0.CO;2-H

Byrne, R. W., & Corp, N. (2004). Neocortex size predicts deception rate in primates. *Proceedings of the Royal Society B – Biological Sciences, 271*(1549), 1693–1699. https://doi.org/10.1098/rspb.2004.2780

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism.* New York: Academic Press.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis, 30,* 1495–1506. https://doi.org/10.1111/j.1539-6924.2010.01460.x

Corry, N., Merritt, R. D., Mrug, S., & Pamp, B. (2008). The factor structure of the Narcissistic Personality Inventory. *Journal of Personality Assessment, 90*(6), 593–600. https://doi.org/10.1080/00223890802388590

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (p. 163–228). Oxford University Press.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R).* Odessa, FL: Psychological Assessment Resources.

Dane, L. K., Jonason, P. K., & McCaffrey, M. (2018). Physiological tests of the cheater hypothesis for the Dark Triad traits: Testosterone, cortisol, and a social stressor. *Personality and Individual Differences, 121*, 227–231. https://doi.org/10.1016/j.paid.2017.09.010

Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society B – Biological Sciences, 205*, 489–511. https://doi.org/10.1098/rspb.1979.0081

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology, 37*(10), 1713–1722. https://doi.org/10.1037/0022-3514.37.10.1713

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Exline, R. V., Thibaut, J., Hickey, C. B., & Gumpert, P. (1970). Visual interaction in relation to Machiavellianism and an unethical act. In R. Christie & F. L. Geis (Eds.), *Studies in Machiavellianism.* New York: Academic Press.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Geis, F. L., & Moon, T. H. (1981). Machiavellianism and deception. *Journal of Personality and Social Psychology, 41*(4), 766–775. https://doi.org/10.1037/0022-3514.41.4.766

Giammarco, E. A., Atkinson, B., Baughman, H. M., Veselka, L., & Vernon, P. A. (2013). The relation between antisocial personality and the perceived ability to deceive. *Personality and Individual Differences, 54*(2), 246–250. https://doi.org/10.1016/j.paid.2012.09.004

Gibney, E. (2016, January). Go players react to computer defeat. *Nature.* https://doi.org/10.1038/nature.2016.19255

Goodfellow, I., Pouget-Abadie, J., Mirza, M.; Xu, B.; Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Proceedings of the International Conference on Neural Information Processing Systems,* 2672–2680.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance? Evidence from AI experts. arXiv preprint, arXiv:1705.08807

Gunnthorsdottir, A., McCabe, K., & Smith, V. (2002). Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *Journal of Economic Psychology, 23*(1), 49–66. https://doi.org/10.1016/S0167-4870(01)00067-8

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology, 53*(1), 7–16. https://doi.org/10.1037/0022-006X.53.1.7

Holtzman, N. S. (2011). Facing a psychopath: Detecting the Dark Triad from emotionally-neutral faces, using prototypes from the personality Faceaurus. *Journal of Research in Personality, 45*(6), 648–654. https://doi.org/10.1016/j.jrp.2011.09.002

Hopwood, C. J., Thomas, K. M., Markon, K. E., Wright, A. G. C., & Krueger, R. F. (2012). DSM-5 personality traits and DSM–IV personality disorders. *Journal of Abnormal Psychology, 121*(2), 424–432. https://doi.org/10.1037/a0026656

Hu, X., Chen, H., & Fu, G. (2012). A repeated lie becomes a truth? The effect of intentional control and training on deception. *Frontiers in Psychology, 3,* Article 488. https://doi.org/10.3389/fpsyg.2012.00488

Jonason, P. K., Foster, J., Oshio, A., Sitnikova, M., Birkas, B., & Gouveia, V. (2017). Self-construals and the Dark Triad traits in six countries. *Personality and Individual Differences, 113*, 120–124. https://doi.org/10.1016/j.paid.2017.02.053

Jonason, P. K., Li, N. P., & Teicher, E. A. (2010). Who is James Bond?: The Dark Triad as an agentic social style*. Individual Differences Research, 8*(2), 111–120.

Jonason, P. K., Lyons, M., Baughman, H. M., & Vernon, P. A. (2014b). What a tangled web we weave: The Dark Triad traits and deception. *Personality and Individual Differences, 70*, 117–119. https://doi.org/10.1016/j.paid.2014.06.038

Jonason, P. K., Slomski, S., & Partyka, J. (2012). The Dark Triad at work: How toxic employees get their way. *Personality and Individual Differences, 52*(3), 449–453. https://doi.org/10.1016/j.paid.2011.11.008

Jonason, P. K., & Webster, G. D. (2012). A protean approach to social influence: Dark Triad personalities and social influence tactics. *Personality and Individual Differences, 52*(4), 521–526. https://doi.org/10.1016/j.paid.2011.11.023

Jonason, P. K., Wee, S., & Li, N. P. (2014a). Thinking bigger and better about “bad apples”: Evolutionary Industrial/Organizational Psychology and the Dark Triad. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 117–121. https://doi.org10.1111/iops.12118

Jones, D. N., & Figueredo, A. J. (2013). The core of darkness: Uncovering the heart of the Dark Triad. *European Journal of Personality, 27*(6), 521–531. https://doi.org/10.1002/per.1893

Jones, D. N., & Paulhus, D. L. (2011). Differentiating the Dark Triad within the interpersonal circumplex. In L. M. Horowitz & S. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (p. 249–267). John Wiley & Sons, Inc..

Jones, D. N., & Paulhus, D. L. (2017). Duplicity among the Dark Triad: Three faces of deceit. *Journal of Personality and Social Psychology, 113*(2), 329–342. https://doi.org/10.1037/pspp0000139

Kashy, D. A., & DePaulo, B. M. (1996). Who lies? *Journal of Personality and Social Psychology, 70*(5), 1037–1051. https://doi.org/10.1037/0022-3514.70.5.1037

Lee, K., & Ashton, M. C. (2005). Psychopathy, Machiavellianism, and narcissism in the five-factor model and the HEXACO model of personality structure. *Personality and Individual Differences, 38*, 1571–1582. https://doi.org/10.1016/j.paid.2004.09.016

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment, 14*(4), 299–316. https://doi.org/10.1111/j.1468-2389.2006.00353.x

Levine, T. R. (2014). Truth-Default Theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communication Monographs, 66*(2), 125–144. https://doi.org/10.1080/03637759909376468

Lowe-Calverley, E., & Grieve, R. (2017). Web of deceit: Relationships between the Dark Triad, perceived ability to deceive and cyberloafing. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 11*(2), Article 5. https://doi.org/10.5817/CP2017-2-5

Lyons, M., Croft, A., Fairhurst, S., Varley, K., & Wilson, C. (2017). Seeing through crocodile tears? Sex-specific associations between the Dark Triad traits and lie detection accuracy. *Personality and Individual Differences, 113*, 1–4. https://doi.org/10.1016/j.paid.2017.03.008

Lyons, M., Healy, N., & Bruno, D. (2013). It takes one to know one: Relationship between lie detection and psychopathy. *Personality and Individual Differences, 55*(6), 676–679. https://doi.org/10.1016/j.paid.2013.05.018

Mealey, L. (1995). The sociobiology of sociopathy: An integrated evolutionary model. *Behavioral and Brain Sciences, 18*(3), 523–599. https://doi.org/10.1017/S0140525X00039595

Monaghan, C., Bizumic, B., & Sellbom, M. (2016). The role of Machiavellian views and tactics in psychopathology. *Personality and Individual Differences, 94,* 72–

81. https://doi.org/10.1016/j.paid.2016.01.002

Monaghan, C., Bizumic, B., & Sellbom, M. (2018). Nomological network of two-dimensional Machiavellianism. *Personality and Individual Differences, 130,* 161–173. https://doi.org/10.1016/j.paid.2018.03.047

Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology, 93*(1), 1–11. https://doi.org/10.1037/0022-3514.93.1.1

Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The malevolent side of human nature: A meta-analysis and critical review of the literature on the Dark Triad (narcissism, Machiavellianism, and psychopathy). *Perspectives on Psychological Science, 12*(2), 183–204. https://doi.org/10.1177/1745691616666070

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (p. 555–572). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-319-26485-1_33

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *American Psychologist, 61*(6), 622–631. https://doi.org/10.1037/0003-066X.61.6.622

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, Volume 171.* Clifton, VA: IOS Press.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36*, 556–563. doi 10.1016/S0092-6566(02)00505-6

Peace, K. A., & Sinclair, S. M. (2012). Cold-blooded lie catchers? An investigation of psychopathy, emotional processing, and deception detection. *Legal and Criminological Psychology, 17*(1), 177–191. https://doi.org/10.1348/135532510X524789

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality, 21*, 549–587. https://doi.org/10.1002/per.629

Penke, L., & Jokela, M. (2016). The evolutionary genetics of personality revisited. *Current Opinion in Psychology, 7*, 104–109. https://doi.org/10.1016/j.copsyc.2015.08.021

Plomin, R. (1986). *Development, genetics, and psychology.* Hillsdale, NJ: Erlbaum

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports, 45*(2), 590. https://doi.org/10.2466/pr0.1979.45.2.590

Rens, N., Schwartenbeck, P., Cunnington, R., & Pezzulo, G. (2020, December 30). Evidence for entropy maximisation in human free choice behaviour. https://doi.org/10.31234/osf.io/3ntku

Roeser, K., McGregor, V. E., Stegmaier, S., Mathew, J., Kübler, A., & Meule, A. (2016). The Dark Triad of personality and unethical behavior at different times of day. *Personality and Individual Differences, 88*, 73–77. https://doi.org/10.1016/j.paid.2015.09.002

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S. Guez, A., ... Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature, 588,* 604–609. https://doi.org/10.1038/s41586-020-

03051-4

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529,* 484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., . . . Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550,* 354–359. https://doi.org/10.1038/nature24270

The Royal Society (2017). *Public views of machine learning: Findings from public research and engagement conducted on behalf of the Royal Society.* Retrieved from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

Vernon, P. A., Villani, V. C., Vickers, L. C., & Harris, J. A. (2008). A behavioral genetic investigation of the Dark Triad and the Big 5. *Personality and Individual Differences, 44*, 445–452. https://doi.org/10.1016/j.paid.2007.09.007

Verschuere, B., & in 't Hout, W. (2016). Psychopathic traits and their relationship with the cognitive costs and compulsive nature of lying in offenders. *PLoS One, 11*(7), Article e0158595. https://doi.org/10.1371/journal.pone.0158595

Verschuere, B., Spruyt, A., Meijer, E. H., & Otgaar, H. (2011). The ease of lying. *Consciousness and Cognition, 20*(2), 908–911. https://doi.org/10.1016/j.concog.2010.10.023

Veselka, L., Schermer, J. A., & Vernon, P. A. (2011). Beyond the Big Five: The Dark Triad and the Supernumerary Personality Inventory. *Twin Research and Human Genetics, 14*, 158–168. http://doi.org/10.1375/twin.14.2.158

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34,* 1–56. https://doi.org/10.1017/S0140525X10001354

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. https://doi.org/10.1016/j.paid.2015.07.025

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology, 22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection. *Current Directions in Psychological Science, 20*(3), 197–200. https://doi.org/10.1177/0963721411409007

Wissing, B. G., & Reinhard, M.-A. (2017). The Dark Triad and the PID-5 maladaptive personality traits: Accuracy, confidence and response bias in judgments of veracity. *Frontiers in Psychology, 8,* Article 1549. https://doi.org/10.3389/fpsyg.2017.01549

Wissing, B. G., & Reinhard, M.-A. (2018). Individual differences in risk perception of artificial intelligence. *Swiss Journal of Psychology, 77*(4), 149–157. https://doi.org/10.1024/1421-0185/a000214

Wissing, B. G., & Reinhard, M.-A. (2019). The Dark Triad and deception perceptions. *Frontiers in Psychology, 10,* Article 1811. https://doi.org/10.3389/fpsyg.2019.01811

Wissner-Gross, A. D., & Freer, C. E. (2013). Causal entropic forces. *Physical Review Letters, 110,* Article 168702. https://doi.org/10.1103/PhysRevLett.110.168702

Wright, G. R. T., Berry, C. J., & Bird, G. (2012). "You can't kid a kidder": association between production and detection of deception in an interactive deception task. *Frontiers in Human Neuroscience, 6,* Article 87. https://doi.org/10.3389/fnhum.2012.00087

Wright, G. R. T., Berry, C. J., Catmur, C., & Bird, G. (2015). Good liars are neither 'dark' nor self-deceptive. *PLoS One, 10*(6), Article e0127315. https://doi.org/10.1371/journal.pone.0127315

Yampolskiy, R. V. (2012). Leakproofing the singularity: Artificial intelligence confinement problem. *Journal of Consciousness Studies, 19,* 194–214.

Yudkowsky, E. S. (2002). *The AI-box experiment.* Retrieved from http://yudkowsky.net/singularity/aibox

Zuckerman, M., DePaulo, B.M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, p. 1–59). New York: Academic Press.

Erklärung Eigenanteil

Universität Kassel, Fachbereich Humanwissenschaften

Erklärung zur kumulativen Dissertation im Promotionsfach Psychologie

Erklärung über den Eigenanteil an den veröffentlichten oder zur Veröffentlichung vorgesehenen wissenschaftlichen Schriften innerhalb meiner Dissertationsschrift, Ergänzung zu § 5a Abs. 4 Satz 1 der Allgemeinen Bestimmungen für Promotionen an der Universität Kassel vom 13. Juni 2011

1. Allgemeine Angaben

Wissing, Benno Gerrit

Institut für Psychologie, Universität Kassel

Thema der Dissertation: „Die Dunkle Triade und Täuschung"

2. Nummerierte Aufstellung der eingereichten Schriften

1. Wissing, B. G., & Reinhard, M.-A. (2017). The Dark Triad and the PID-5 maladaptive personality traits: Accuracy, confidence and response bias in judgments of veracity. *Frontiers in Psychology, 8,* Article 1549. https://doi.org/10.3389/fpsyg.2017.01549
2. Wissing, B. G., & Reinhard, M.-A. (2018). Individual differences in risk perception of artificial intelligence. *Swiss Journal of Psychology, 77*(4), 149–157. https://doi.org/10.1024/1421-0185/a000214
3. Wissing, B. G., & Reinhard, M.-A. (2019). The Dark Triad and deception perceptions. *Frontiers in Psychology, 10,* Article 1811. https://doi.org/10.3389/fpsyg.2019.01811

3. Darlegung des eigenen Anteils an diesen Schriften

Zu Nr. 1:

- Konzeptionsentwicklung: überwiegend
- Literaturrecherche: vollständig
- Methodenentwicklung: überwiegend
- Versuchsdesignentwicklung: überwiegend
- Datenerhebung: vollständig
- Datenauswertung: vollständig
- Ergebnisdiskussion: vollständig
- Manuskripterstellung: überwiegend
- Review-Prozess-Bewältigung: überwiegend

Zu Nr. 2:

- Konzeptionsentwicklung: vollständig
- Literaturrecherche: vollständig
- Methodenentwicklung: vollständig
- Versuchsdesignentwicklung: vollständig
- Datenerhebung: überwiegend
- Datenauswertung: vollständig
- Ergebnisdiskussion: vollständig
- Manuskripterstellung: vollständig
- Review-Prozess-Bewältigung: überwiegend

Zu Nr. 3:

- Konzeptionsentwicklung: teilweise
- Literaturrecherche: überwiegend

- Methodenentwicklung: teilweise
- Versuchsdesignentwicklung: teilweise
- Datenerhebung: vollständig
- Datenauswertung: vollständig
- Ergebnisdiskussion: vollständig
- Manuskripterstellung: vollständig
- Review-Prozess-Bewältigung: überwiegend

4. Anschriften der Mitautoren

Marc-André Reinhard: reinhard@psychologie.uni-kassel.de

…………………………………………………………

Kassel, den 01.01.2021 Benno Gerrit Wissing

Ich bestätige die von Herrn Wissing unter Pkt. 3 abgegebene Erklärung:

1.

Prof. Dr. Marc-André Reinhard Unterschrift:

1.1.2021
……………………………

Anhang

BRIEF RESEARCH REPORT
published: 13 August 2019
doi: 10.3389/fpsyg.2019.01811

The Dark Triad and Deception Perceptions

Benno Gerrit Wissing[1]* *and Marc-André Reinhard*[2]

[1]*Department of Psychology, University of Kassel, Kassel, Germany,* [2]*Department of Psychology, Social Psychology, University of Kassel, Kassel, Germany*

The present cross-sectional study (*N* = 205) tested the hypothesis that the Dark Triad traits – narcissism, Machiavellianism, and psychopathy – and the PID-5 maladaptive personality traits – Negative Affectivity, Detachment, Antagonism, Disinhibition, and Psychoticism – are associated with specific deception-related perceptions: perceived cue-based deception detectability, perceived deception production ability, and perceived deception detection ability. Participants completed personality and deception measures in an online setting. All three Dark Triad traits and Antagonism were associated with perceived deception production ability, but not (substantially) with perceived deception detection ability and cue-based deception detectability. The results provide a more fine-grained picture of biases associated with the Dark Triad traits in the context of deception and further support the relevance of Antagonism and Detachment as deception-relevant personality traits.

Keywords: dark triad, pid-5, deception, deception detection, deception production

OPEN ACCESS

Edited by:
Renate Volbert,
Psychologische Hochschule Berlin,
Germany

Reviewed by:
Peter Karl Jonason,
Western Sydney University, Australia
Matthias Ziegler,
Humboldt University of Berlin,
Germany

****Correspondence:***
Benno Gerrit Wissing
wissing@uni-kassel.de

Specialty section:
This article was submitted to
Forensic and Legal Psychology,
a section of the journal
Frontiers in Psychology

Received: *29 April 2019*
Accepted: *22 July 2019*
Published: *13 August 2019*

Citation:
Wissing BG and Reinhard M-A (2019)
The Dark Triad and
Deception Perceptions.
Front. Psychol. 10:1811.
doi: 10.3389/fpsyg.2019.01811

INTRODUCTION

Research on the Dark Triad (Paulhus and Williams, 2002) – the moderately associated personality traits of narcissism, Machiavellianism, and psychopathy – has accrued in the last 17 years. Viewed through the lens of the two main personality frameworks, the Five-Factor Model (FFM; Costa and McCrae, 1992) and the HEXACO model (Lee and Ashton, 2005), the three traits are constituted primarily by low levels of agreeableness (Paulhus and Williams, 2002) and low Honesty-Humility (Lee and Ashton, 2005). Given that agreeableness can be defined as the willingness to cooperate, a low manifestation of this trait can have negative consequences for group inclusion (Buss, 1991). Accordingly, individuals high in Dark Triad traits tend to be lower in communion and more agentic (Jonason et al., 2010; Jones and Paulhus, 2010) and Machiavellianism and narcissism are linked with higher independent self-construals (Jonason et al., 2017). Beyond their shared variance, the Dark Triad traits also exhibit substantial distinctiveness: Narcissism is defined by entitlement, grandiosity, dominance, and superiority (Raskin and Hall, 1979; Corry et al., 2008). Machiavellianism is characterized by an amoral, cold, and cynical view of the world and strategic manipulativeness (Christie and Geis, 1970). Psychopathy is associated with higher risk-taking, impulsivity, lower empathy, and lower neuroticism (Hare, 1985). Some researchers have argued for domain-specific adaptiveness of the Dark Triad traits (see Jonason et al., 2014a).

According to a foundational meta-analysis, humans are slightly better than chance at detecting deception (Bond and DePaulo, 2006). Moreover, humans tend to be truth-biased in their response, that is, they assume that others are truthful independently of their actual truth status (Levine et al., 1999) and more variance is found in response bias than in deception detection accuracy (Bond and DePaulo, 2008; Schindler and Reinhard, 2015). The almost chance converging deception detection performance might be explained by the low availability of behavioral cues to deception (Hartwig and Bond, 2011).

Low agreeableness and low Honesty-Humility – the primary correlates of the Dark Triad traits within well-established personality frameworks – are both associated with deceptive behavior. Correspondingly, one attempt to capture the personality-theoretical core of the three traits has converged on manipulation-callousness (Jones and Figueredo, 2013) – thus, further highlighting the close connection between the Dark Triad traits and deception. There is evidence that especially Machiavellianism and psychopathy are associated with interpersonal deception production frequency (Kashy and DePaulo, 1996; Baughman et al., 2014; Jonason et al., 2014b) and the former additionally with amplitude (high-stakes deception; Azizli et al., 2016). On the contrary, narcissism is associated with intrapersonal deception, that is, self-deception (for example, Paulhus and Williams, 2002). Three contextual variables have been identified that influence deceptive behavior exhibited by the Dark Triad traits: risk level, ego depletion, and deception target (self vs. others) (Jones and Paulhus, 2017). A study which measured the Dark Triad traits and the levels of testosterone and cortisol before and after a deception production task (video-taped lying) found a physiological response pattern – cortisol decrease for men high in Machiavellianism (and suggestively so for narcissism and psychopathy as well) and testosterone increase post-test for narcissism and psychopathy – in line with the hypothesis that the Dark Triad traits represent an evolved "cheater strategy" (Dane et al., 2018).

Research on the relationship between deception detection accuracy and personality traits is limited (Aamodt and Custer, 2006). The data in the existing literature on dark personality traits and deception detection ability are mixed. For example, primary psychopathy is associated with lie detection accuracy in men (Lyons et al., 2013), while other research did not report such an association (for example, Peace and Sinclair, 2012). Additionally, no superior lie detection accuracy for individuals high in Machiavellianism was reported (Zuckerman et al., 1981), except for one study that reported woman high in Machiavellianism displayed a higher lie detection ability (Lyons et al., 2017). Moreover, in an interactive deception task, no correlation between deception detection or deception production ability and the Dark Triad traits was detected (Wright et al., 2015). Parallel results regarding deception detection ability were reported (Wissing and Reinhard, 2017). This line of research corresponds to the result pattern of previous research (Bond and DePaulo, 2006, 2008) and expands it to the Dark Triad traits, indicating that these traits are not substantially associated with actual deception detection or deception production ability, but with varying amplitudes and frequencies of deception production and various deception-related judgmental and perceptual biases.

In sum, previous research suggests that self-perceived deception production abilities (Giammarco et al., 2013) and multiple deception-related biases (Wissing and Reinhard, 2017) are associated with the Dark Triad traits. Based on these findings, this study tried to replicate the association of Dark Triad traits with self-perceived deception production abilities and additionally investigated if the Dark Triad personality traits are also correlated with perceptions potentially relevant to the process of deception detection. In detail and beyond the replication regarding self-perceived deception production abilities, this study assumed that all three Dark Triad traits – narcissism, Machiavellianism, and psychopathy – are positively correlated with perceived cue-based deception detectability and self-perceived deception detection abilities. Based on the finding that the Big Five trait Agreeableness is moderately negatively correlated with perceived ability to deceive (Giammarco et al., 2013), it was assumed that the corresponding dimension in PID-5 maladaptive personality space – Antagonism – is positively correlated with self-perceived deception detection abilities. The PID-5 maladaptive personality traits were also included instead of the Big Five, because research found that the PID-5 traits outperform the Big Five as predictors of the Dark Triad traits (Grigoras and Wille, 2017).

MATERIALS AND METHODS

The study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) and the American Psychological Association (APA). Moreover, by the time the data were acquired, it was also not required at Kassel University, nor at most other German universities to seek ethics approval for simple studies on personality and attitudes. Thus, ethics approval was not required as per applicable institutional and national guidelines and regulations. The study exclusively makes use of anonymous questionnaires. No identifying information was obtained from participants. The participants were explicitly informed that the data are treated confidentially. The informed consent of the participants was provided online: every participant had to agree to the following statements: "I understand that my participation is voluntary and that I may withdraw from the study at any time without explanation" and "I hereby confirm that I am at least 18 years old, and that I agree to take part in this study." Furthermore, they could withdraw from the study at any time.

STATISTICAL POWER AND PARTICIPANTS

The correlations between the Dark Triad traits were estimated based on the average effect sizes (*r*) for the intercorrelations between the three traits reported in a meta-analysis

(rs = 0.34, 0.38, 0.58; Muris et al., 2017). The correlations between the three Dark Triad traits and cue-based deception detectability and perceived deception detection ability were estimated as r = 0.25. The r = 0.25 was selected based on an assumed weaker correlation of the Dark Triad traits with cue-based deception detectability and perceived deception detection ability vs. perceived deception production ability (r's = 0.33–0.41; Giammarco et al., 2013). These values were entered into the statistical power analysis tool G*Power (Faul et al., 2009) and the required minimum sample size N = 157 for a linear multiple regression with k = 3 predictors at alpha level α = 0.05 with Power 1 − β = 0.95 for an estimated small to medium effect size f^2 = 0.11 (calculated via p^2) was calculated. For a simple correlation test of r = 0.25 (two-tailed) at alpha level α = 0.05 with Power 1 − β = 0.95, the required sample size was N = 202.

Participants were recruited with Amazon's Mechanical Turk (MTurk), selecting exclusively MTurk Masters (a high-performance group that demonstrated accuracy in the past per Amazon), and were paid a small fee (1$). n = 17 participants dropped out before the first dependent variable measurement was finished, resulting in the final sample (N = 205; 58.54% male, 41.46% female; M_{age} = 37.76, SD = 10.80, age range: 22–70 years; 94.8% native English speaker, 93.66% living in the United States of America, 73.66% Caucasian; 72.68% employees).

PROCEDURE AND MEASURES

The Dark Triad traits were measured with the Short Dark Triad (SD3; Jones and Paulhus, 2014), a 27-item short self-report instrument that measures narcissism (for example, "People see me as a natural leader."; α = 0.87), Machiavellianism (for example, "I like to use clever manipulation to get my way."; α = 0.87), and psychopathy (for example, "People who mess with me always regret it."; α = 0.81) with nine items each on a 5-point Likert-type scale (1 = *disagree strongly*, 5 = *agree strongly*).

The Personality Inventory for DSM-5 Brief Form (PID-5-BF; Krueger et al., 2012; American Psychiatric Association, 2013) was used to measure the five maladaptive personality trait domains of the PID-5 model with 25 items in total, consisting of Negative Affectivity (for example, "I worry about almost everything"; α = 0.82), Detachment (for example, "I often feel like nothing I do really matters"; α = 0.83), Antagonism (for example, "It's no big deal if I hurt other peoples' feelings"; α = 0.83), Disinhibition (for example, "People would describe me as reckless"; α = 0.88), and Psychoticism (for example, "My thoughts often don't make sense to others"; α = 0.85) using a 4-point Likert-type scale (0 = *very false/often false*, 3 = *very true/often true*).

Participants read that the next questions would be about their thoughts on how people behave when they are lying or telling the truth. Cue-based deception detectability was measured using 22 statements based on previously identified beliefs about cues of deception (Hartwig and Bond, 2011) referring to verbal (three items; for example, "Deceptive statements are less detailed than truthful statements"), para-verbal (four items; for example, "Liars pause less than truth tellers"), and non-verbal aspects (15 items; "Liars blink less than truth tellers") of deceptive behavior. A 7-point Likert-type scale ranging from −3 (for example, "Deceptive statements are less detailed than truthful statements") *via* 0 (= *no difference*) to +3 (for example, "Deceptive statements are more detailed than truthful statements") was used. Negative values were recoded to positive ones (−1 = 1, −2 = 2, −3 = 3), resulting in the final scale [value range: (0, 3)][1]. (Note 1) For each statement, participants should select the answer that was most closely aligned with their opinion on the given statement. Exploratory factor analysis using maximum likelihood suggested a one-factor structure as being sufficient with factor loadings between 0.42 and 0.79 (40% explained variance). One item with factor loading = 0.42 < 0.50 was removed, resulting in the final scale with 21 items. Exploratory factor analysis on the final scale using maximum likelihood suggested a one-factor structure as being sufficient with factor loadings between 0.55 and 0.80 (41% explained variance) (α = 0.93, ω_h = 0.76, ω_t = 0.95).

Using a 7-point Likert-type scale (1 = *strongly disagree*, 7 = *strongly agree*) perceived deception detection ability was measured with six items (for example, "In general, liars can be easily recognized," "In general, I'm good at detecting liars"; α = 0.85). Exploratory factor analysis using maximum likelihood suggested a one-factor structure as being sufficient with factor loadings between 0.36 and 0.96 (51% explained variance). One item with factor loading = 0.36 < 0.50 was removed, resulting in the final scale with five items. Exploratory factor analysis on the final scale using maximum likelihood suggested a one-factor structure as being sufficient with factor loadings between 0.63 and 0.98 (58% explained variance) (α = 0.86, ω_h = 0.72, ω_t = 0.93).

Using a 7-point Likert-type scale (1 = *strongly disagree*, 7 = *strongly agree*) perceived deception production ability was assessed with three items (for example, "I'm a good liar"; α = 0.77, ω_t = 0.78). Exploratory factor analysis using maximum likelihood suggested a one-factor structure as being sufficient with factor loadings between 0.71 and 0.76 (53% explained variance).

RESULTS

Descriptive statistics for all variables can be seen in **Table 1**.

Cue-Based Deception Detectability

As seen in **Table 2**, there was no substantial association pattern between dark and maladaptive personality traits and

[1] This recoding step is necessary, because (1) perceived low and high cue-frequency/-amplitude under deception relative to no deception can make a cue deception detection relevant, and (2) without the recoding step, cues with perceived low vs. high cue-frequency/-amplitude under deception relative to no deception would cancel each other out.

TABLE 1 | Descriptive statistics.

	***M* (SD)**		
	Overall (*N* = 205)	**Sex: male (*n* = 120)**	**Sex: female (*n* = 85)**
Deception			
Cue-based deception detectability	1.42 (0.65)	1.42 (0.57)	1.42 (0.75)
Perceived deception detection ability	4.40 (1.16)	4.54 (1.05)	4.20 (1.28)
Perceived deception production ability	3.64 (1.36)	3.83 (1.36)	3.39 (1.33)
Dark Triad			
Narcissism	2.54 (0.81)	2.77 (0.85)	2.22 (0.63)
Machiavellianism	3.00 (0.79)	3.19 (0.83)	2.74 (0.65)
Psychopathy	2.12 (0.73)	2.31 (0.72)	1.84 (0.65)
PID-5			
Negative affectivity	0.94 (0.71)	0.85 (0.68)	1.06 (0.74)
Detachment	0.88 (0.75)	0.92 (0.76)	0.83 (0.73)
Antagonism	0.61 (0.61)	0.72 (0.65)	0.47 (0.54)
Disinhibition	0.54 (0.64)	0.57 (0.62)	0.49 (0.66)
Psychoticism	0.57 (0.63)	0.61 (0.58)	0.52 (0.69)

TABLE 2 | Zero-order Pearson correlation coefficients with 95% CIs (in brackets) for the Dark Triad traits, the PID-5 traits and the deception variables.

	Cue-based deception detectability	**Perceived deception detection ability**	**Perceived deception production ability**
Dark Triad			
Narcissism	−0.01 [−0.15, 0.13]	0.16 [0.02, 0.29]*	0.33 [0.21, 0.45]***
Machiavellianism	−0.05 [−0.18, 0.09]	0.12 [−0.02, 0.25]	0.45 [0.34, 0.56]***
Psychopathy	−0.01 [−0.14, 0.13]	0.14 [0.01, 0.28]*	0.44 [0.32, 0.54]***
PID-5			
Negative affectivity	0.02 [−0.12, 0.15]	−0.13 [−0.26, 0.01]	0.03 [−0.11, 0.17]
Detachment	−0.13 [−0.27, 0.00]	−0.17 [−0.30, −0.04]*	0.02 [−0.12, 0.16]
Antagonism	−0.03 [−0.17, 0.10]	0.08 [−0.06, 0.21]	0.45 [0.34, 0.56]***
Disinhibition	−0.01 [−0.14, 0.13]	0.03 [−0.10, 0.17]	0.20 [0.06, 0.33]**
Psychoticism	0.04 [−0.10, 0.17]	−0.01 [−0.15, 0.12]	0.15 [0.01, 0.28]*

*N = 205; *p < 0.05, **p < 0.01, ***p < 0.001 (two-tailed).*

cue-based deception detectability. Only Detachment showed a suggestive negative association pattern with cue-based deception detectability.

A linear regression with cue-based deception detectability as the criterion and the Dark Triad traits as predictors produced a statistically non-significant model [adjusted R^2 = −0.01, F(3, 201) = 0.23, p = 0.88]. A hierarchical multiple regression analysis with cue-based deception detectability as the criterion variable, the PID-5 traits entered in Step 1, was stopped after Step 1 [adjusted R^2 = 0.02, F(5, 199) = 1.77, p = 0.12].

Perceived Deception Detection Ability

As seen in **Table 2**, among the Dark Triad traits, narcissism and psychopathy were weakly associated with perceived deception detection ability. The pattern for Machiavellianism was suggestive of a potential association. Among the PID-5 traits, Detachment was negatively associated with perceived deception detection ability.

A linear regression with perceived deception detection ability as the criterion and the Dark Triad traits as predictors produced a statistically non-significant model [adjusted R^2 = 0.01, F(3, 201) = 2.01, p = 0.11]. A hierarchical multiple regression analysis with perceived deception detection ability as the criterion variable, the PID-5 traits entered in Step 1, adjusted R^2 = 0.03, F(5, 199) = 2.42, p = 0.04, and the Dark Triad traits entered in Step 2, adjusted R^2 = 0.03, F(8, 196) = 1.92, p = 0.06, as predictor variables was computed. In the first step, Detachment emerged as a substantial predictor [b = −0.30, 95% CI = (−0.55, −0.04), p = 0.02]. The Dark Triad traits did not account for additional variance in perceived deception detection ability above the PID-5 traits, ΔR^2 = 0.00, F(3, 196) = 1.09, p = 0.35.

Perceived Deception Production Ability

As seen in **Table 2**, all three Dark Triad traits were correlated with perceived deception production ability. Among the PID-5 traits, Antagonism, Disinhibition, and Psychoticism were associated with perceived deception production ability.

In a linear regression with perceived deception production ability as the criterion and the Dark Triad traits as predictors [adjusted R^2 = 0.23, F(3, 201) = 21.43, p < 0.001], Machiavellianism [b = 0.48, 95% CI = (0.20, 0.76), p < 0.001] and psychopathy [b = 0.42, 95% CI = (0.10, 0.73), p = 0.01] emerged as substantial predictors. A hierarchical multiple regression analysis with perceived deception production ability as the criterion variable, the PID-5 traits entered as predictor variables in Step 1, adjusted R^2 = 0.22, F(5, 199) = 12.46, p < 0.001, and the Dark Triad traits entered in Step 2, adjusted R^2 = 0.25, F(8, 196) = 9.56, p < 0.001, was computed. In the first step, Antagonism emerged as a substantial predictor [b = 1.32, 95% CI = (0.95, 1.68), p < 0.001]. The Dark Triad traits accounted for additional variance in perceived deception production ability above the PID-5 traits, ΔR^2 = 0.03, F(3, 196) = 3.84, p = 0.01. In Step 2, Antagonism remained a substantial predictor [b = 0.34, 95% CI = (0.11, 0.58), p = 0.004] and Machiavellianism emerged as an additional predictor [b = 0.23, 95% CI = (0.05, 0.41), p = 0.01].

DISCUSSION

The data indicate that the Dark Triad traits and Antagonism are associated with perceived deception production ability, but not (substantially) with perceived deception detection ability and cue-based deception detectability. This replicates the result pattern regarding the Dark Triad traits and regarding Antagonism [r(204) = 0.45], which is the maladaptive form of Big Five Agreeableness [r(1447) = −0.28; Giammarco et al., 2013]. Also, a comparable adjusted R^2 = 0.23 was found for the Dark Triad traits as predictors of perceived ability to deceive [adjusted R^2s = 0.12–0.22; Giammarco et al., 2013].

Beyond Antagonism and among the PID-5 maladaptive personality traits, Disinhibition and Psychoticism were associated with perceived deception production ability. Controlling for the shared variance in a regression model, only Antagonism remained a substantial predictor of perceived deception production ability. Interestingly, when the Dark Triad traits and the PID-5 traits as a whole – PID-5 Antagonism is one of the proposed origins of the shared variance of the Dark Triad – were used as predictor variables of perceived deception production ability, only Machiavellianism remained a substantial predictor among the Dark Triad traits. This result suggests that Machiavellianism can uniquely explain variance above the shared variance of the Dark Triad traits and PID-5 maladaptive personality traits in perceived deception production ability. This contrasts with a meta-analysis on the Dark Triad that suggested that "psychopathy runs the show" regarding psychosocial correlates of the Dark Triad (Muris et al., 2017), suggesting that at least in the psychosocial relevant domain of deception, Machiavellianism might play an important role. Given that Machiavellianism is associated with deception production frequency (Kashy and DePaulo, 1996; Baughman et al., 2014; Jonason et al., 2014b) and amplitude (high-stakes deception; Azizli et al., 2016), a relatively high perceived deception production ability might be a necessary precondition or a consequence (or both) of this high-frequency, high-amplitude deceptive behavioral pattern.

Detachment was negatively associated with perceived deception detection ability and suggestively with cue-based deception detectability, which might be explained partially by the central interpersonally outcome of Detachment, which is withdrawal from other people. This social withdrawal might lead individuals high in Detachment to regard their deception detection skills as being relatively low, given the relative absence of social interaction as a precondition for potential deception ability acquisition and evaluation. This finding further supports the relevance of Detachment in the context of deception, for example, previously, Detachment has been found to be negatively correlated with response bias in deception detection (Wissing and Reinhard, 2017).

All three Dark Triad traits are multi-dimensional constructs: narcissism can be differentiated in grandiose vs. vulnerable expressions (Miller and Campbell, 2008; Pincus and Lukowitsky, 2010) and is conceptualized as consisting of three higher order factors (Antagonism, Neuroticism, and Agentic Extraversion) (Miller et al., 2016); psychopathy consists of primary and secondary forms (Levenson et al., 1995); Machiavellianism can be defined as a two-dimensional construct (tactics vs. views; Monaghan et al., 2016). The used SD3 measure of the Dark Triad traits tends to capture secondary psychopathy and grandiose narcissism (Jones and Paulhus, 2014). The distinction between Machiavellianism and psychopathy of the SD3 is based on impulsivity (low for Machiavellianism, high for psychopathy; see also Jones and Paulhus (2009)), but this difference in impulsivity has been questioned (Miller et al., 2017). The conceptual distinctiveness of the Dark Triad traits has also been questioned by a meta-analysis (Muris et al., 2017). Given the uncaptured multi-dimensionality and partially questionable validity of the constructs, the results of the present study might be limited and should be resolved by future studies. Additionally, it is worth stressing that self-rated ability and actual ability are not necessarily associated and might be orthogonal or even negatively correlated.

While the usage of MTurk samples is debatable, there is evidence that they are comparable to laboratory scenarios (Thomas and Clifford, 2017). The present study was based entirely on self-report instruments, suggesting that the reported results might be subject to method bias. Effect sizes were relatively low, with the exception of self-perceived deception production ability. The used deception scales were self-constructed and not properly validated. While exploratory factor analysis, internal consistency, and general factor saturation indices suggested good scale properties – and the perceived deception production ability scale showed converged validity with the PATD scale (Schneider and Goffin, 2012) in terms of close intercorrelation point estimates with the Dark Triad traits – the results should be replicated with properly validated measures by future studies before building upon them in other ways, for example, future studies might try to replicate the result pattern of this study regarding the Dark Triad traits and Antagonism as predictors of perceived deception production ability with the PATD scale (Schneider and Goffin, 2012) as the criterion variable. Also, studies with more statistical power should be conducted to investigate the potential association between narcissism and psychopathy with perceived deception detection ability. Future studies might also explore the ecological validity of the deception perceptions – when properly validated measures exist – by testing if these are predictive of actual deceptive behavior.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) and the American Psychological Association (APA). Moreover, by the time the data were acquired, it was also not required at Kassel University, nor at most other German universities to seek ethics approval for simple studies on personality and attitudes. The study exclusively makes use of anonymous questionnaires. No identifying information was obtained from participants. The participants were explicitly informed that the data are treated confidentially. Every participant had to agree to the following statements: "I understand that my participation is voluntary and that I may withdraw from the study at any time without explanation" and "I hereby confirm that I am at least 18 years old, and that I agree to take part in this study." Furthermore, they could withdraw from the study at any time.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Aamodt, M. G., and Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Exam.* 15, 6–11.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Arlington, VA: American Psychiatric Association.

Azizli, N., Atkinson, B. E., Baughman, H. M., Chin, K., Vernon, P. A., Harris, E., et al. (2016). Lies and crimes: Dark Triad, misconduct, and high-stakes deception. *Personal. Individ. Differ.* 89, 34–39. doi: 10.1016/j.paid.2015.09.034

Baughman, H. M., Jonason, P. K., Lyons, M., and Vernon, P. A. (2014). Liar liar pants on fire: cheater strategies linked to the Dark Triad. *Personal. Individ. Differ.* 71, 35–38. doi: 10.1016/j.paid.2014.07.019

Bond, C. F., and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personal. Soc. Psychol. Rev.* 10, 214–234. doi: 10.1207/s15327957pspr1003_2

Bond, C. F., and DePaulo, B. M. (2008). Individual differences in judging deception: accuracy and bias. *Psychol. Bull.* 134, 477–492. doi: 10.1037/0033-2909.134.4.477

Buss, D. M. (1991). Evolutionary personality psychology. *Annu. Rev. Psychol.* 42, 459–491. doi: 10.1146/annurev.ps.42.020191.002331

Christie, R., and Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Corry, N., Merritt, R. D., Mrug, S., and Pamp, B. (2008). The factor structure of the narcissistic personality inventory. *J. Pers. Assess.* 90, 593–600. doi: 10.1080/00223890802388590

Costa, P. T., and McCrae, R. R. (1992). *NEO five-factor inventory*. Lutz, FL: Psychological Assessment Resources.

Dane, L. K., Jonason, P. K., and Walker, M. (2018). The hormones of a cheater: the Dark Triad traits, testosterone, cortisol, and stress. *Personal. Individ. Differ.* 121, 227–231. doi: 10.1016/j.paid.2017.09.010

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Giammarco, E. A., Atkinson, B., Baughman, H., Veselka, L., and Vernon, P. A. (2013). The relation between antisocial personality and the perceived ability to deceive. *Personal. Individ. Differ.* 54, 246–250. doi: 10.1016/j.paid.2012.09.004

Grigoras, M., and Wille, B. (2017). Shedding light on the dark side: associations between the dark triad and the DSM-5 maladaptive trait model. *Personal. Individ. Differ.* 104, 516–521. doi: 10.1016/j.paid.2016.09.016

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *J. Consult. Clin. Psychol.* 53, 7–16. doi: 10.1037/0022-006X.53.1.7

Hartwig, M., and Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychol. Bull.* 137, 643–659. doi: 10.1037/a0023589

Jonason, P. K., Foster, J. D., Oshio, A., Sitnikova, M., Birkas, B., and Gouveia, V. V. (2017). Self-construals and the Dark Triad traits in six countries. *Personal. Individ. Differ.* 113, 120–124. doi: 10.1016/j.paid.2017.02.053

Jonason, P. K., Li, N. P., and Teicher, E. A. (2010). Who is James Bond? The Dark Triad as an agentic social style. *Individ. Differ. Res.* 8, 111–120.

Jonason, P. K., Lyons, M., Baughman, H. M., and Vernon, P. A. (2014b). What a tangled web we weave: the Dark Triad and deception. *Personal. Individ. Differ.* 70, 117–119. doi: 10.1016/j.paid.2014.06.038

Jonason, P. K., Wee, S., and Li, N. P. (2014a). Thinking bigger and better about "bad apples": evolutionary industrial-organizational psychology and the Dark Triad. *Ind. Organ. Psychol.* 7, 117–121. doi: 10.1111/iops.12118

Jones, D. N., and Figueredo, A. J. (2013). The core of darkness: uncovering the heart of the Dark Triad. *Eur. J. Personal.* 27, 521–531. doi: 10.1002/per.1893

Jones, D. N., and Paulhus, D. L. (2009). "Machiavellianism" in *Handbook of individual differences in social behavior*. eds. M. R. Leary and R. H. Hoyle (New York, NY: Guilford), 93–108.

Jones, D. N., and Paulhus, D. L. (2010). "Differentiating the Dark Triad within the interpersonal circumplex" in *Handbook of interpersonal theory and research*. eds. L. M. Horowitz and S. N. Strack (New York: Guilford), 249–267.

Jones, D. N., and Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): a brief measure of dark personality traits. *Assessment* 21, 28–41. doi: 10.1177/1073191113514105

Jones, D. N., and Paulhus, D. L. (2017). Duplicity among the dark triad: three faces of deceit. *J. Pers. Soc. Psychol.* 113, 329–342. doi: 10.1037/pspp0000139

Kashy, D. A., and DePaulo, B. M. (1996). Who lies? *J. Pers. Soc. Psychol.* 70, 1037–1051. doi: 10.1037/0022-3514.70.5.1037

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., and Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychol. Med.* 42, 1879–1890. doi: 10.1017/S0033291711002674

Lee, K., and Ashton, M. C. (2005). Psychopathy, Machiavellianism, and Narcissism in the dive-factor model and the HEXACO model of personality structure. *Personal. Individ. Differ.* 38, 1571–1582. doi: 10.1016/j.paid.2004.09.016

Levenson, M. R., Kiehl, K. A., and Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *J. Pers. Soc. Psychol.* 68, 151–158. doi: 10.1037/0022-3514.68.1.151

Levine, T. R., Park, H. S., and McCornack, S. A. (1999). Accuracy in detecting truths and lies: documenting the "veracity effect". *Commun. Monogr.* 66, 125–144. doi: 10.1080/03637759909376468

Lyons, M., Croft, A., Fairhurst, S., Varley, K., and Wilson, C. (2017). Seeing through crocodile tears? Sex-specific associations between the Dark Triad traits and lie detection accuracy. *Personal. Individ. Differ.* 113, 1–4. doi: 10.1016/j.paid.2017.03.008

Lyons, M., Healy, N., and Bruno, D. (2013). It takes one to know one – psychopathy and deception detection ability. *Personal. Individ. Differ.* 55, 676–679. doi: 10.1016/j.paid.2013.05.018

Miller, J. D., and Campbell, W. K. (2008). Comparing clinical and social-personality conceptualizations of narcissism. *J. Pers.* 76, 449–476. doi: 10.1111/j.1467-6494.2008.00492.x

Miller, J. D., Hyatt, C. S., Maples-Keller, J. L., Carter, N. T., and Lynam, D. R. (2017). Psychopathy and Machiavellianism: a distinction without a difference? *J. Pers.* 85, 439–453. doi: 10.1111/jopy.12251

Miller, J. D., Lynam, D. R., McCain, J. L., Few, L. R., Crego, C., Widiger, T. A., et al. (2016). Thinking structurally about narcissism: an examination of the five-factor narcissism inventory and its components. *J. Personal. Disord.* 30, 1–18. doi: 10.1521/pedi_2015_29_177

Monaghan, C., Bizumic, B., and Sellbom, M. (2016). The role of Machiavellian views and tactics in psychopathology. *Personal. Individ. Differ.* 94, 72–81. doi: 10.1016/j.paid.2016.01.002

Muris, P., Merckelbach, H., Otgaar, H., and Meijer, E. (2017). The malevolent side of human nature: a meta-analysis and critical review of the literature on the Dark Triad (narcissism, Machiavellianism, and psychopathy). *Perspect. Psychol. Sci.* 12, 183–204. doi: 10.1177/1745691616666070

Paulhus, D. L., and Williams, K. M. (2002). The Dark Triad of personality: narcissism, Machiavellianism, and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/S0092-6566(02)00505-6

Peace, K. A., and Sinclair, S. M. (2012). Cold-blooded lie catchers? An investigation of psychopathy, emotional processing, and deception detection. *Leg. Criminol. Psychol.* 17, 177–191. doi: 10.1348/135532510X524789

Pincus, A. L., and Lukowitsky, M. R. (2010). Pathological narcissism and narcissistic personality disorder. *Annu. Rev. Clin. Psychol.* 6, 421–446. doi: 10.1146/annurev.clinpsy.121208.131215

Raskin, R., and Hall, C. S. (1979). A Narcissistic Personality Inventory. *Psychol. Rep.* 45:590. doi: 10.2466/pr0.1979.45.2.590

Schindler, S., and Reinhard, M.-A. (2015). Increasing skepticism toward potential liars: effects of existential threat on veracity judgements and the moderating role of honesty norm activation. *Front. Psychol.* 6:1312. doi: 10.3389/fpsyg.2015.01312

Schneider, T. J., and Goffin, R. D. (2012). Perceived ability to deceive and incremental prediction in pre-employment personality testing. *Personal. Individ. Differ.* 52, 806–811. doi: 10.1016/j.paid.2012.01.015

Thomas, K. A., and Clifford, S. (2017). Validity and mechanical turk: an assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.* 77, 184–197. doi: 10.1016/j.chb.2017.08.038

Wissing, B. G., and Reinhard, M.-A. (2017). The Dark Triad and the PID-5 maladaptive personality traits: accuracy, confidence and response bias in judgments of veracity. *Front. Psychol.* 8:1549. doi: 10.3389/fpsyg.2017.01549

Wright, G. R. T., Berry, C. J., Catmur, C., and Bird, G. (2015). Good liars are neither 'dark' nor self-deceptive. *PLoS One* 10:e0127315. doi: 10.1371/journal.pone.0127315

Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Adv. Exp. Soc. Psychol.* 14, 1–59. doi: 10.1016/S0065-2601(08)60369-X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 September 2017
doi: 10.3389/fpsyg.2017.01549

The Dark Triad and the PID-5 Maladaptive Personality Traits: Accuracy, Confidence and Response Bias in Judgments of Veracity

Benno G. Wissing[1]* *and Marc-André Reinhard*[2]

[1] Department of Psychology, Cognitive Psychology, University of Kassel, Kassel, Germany, [2] Department of Psychology, Social Psychology, University of Kassel, Kassel, Germany

The Dark Triad traits—narcissism, Machiavellianism and psychopathy—have been found to be associated with intra- or interpersonal deception production frequency. This cross-sectional study ($N = 207$) investigated if the Dark Triad traits are also associated with deception detection accuracy, as implicated by the recent conception of a deception-general ability. To investigate associations between maladaptive personality space and deception, the PID-5 maladaptive personality traits were included to investigate if besides Machiavellianism, Detachment is negatively associated with response bias. Finally, associations between the Dark Triad traits, Antagonism, Negative Affectivity and confidence judgments were investigated. Participants watched videos of lying vs. truth-telling senders and judged the truthfulness of the statements. None of the Dark Triad traits was found to be associated with the ability to detect deception. Detachment was negatively associated with response bias. Psychopathy was associated with global confidence judgments. The results provide additional support that dark and maladaptive personality traits are associated with judgmental biases but not with accuracy in deception detection. The internal consistencies of 4 of the 8 subscales of the used personality short scales were only low and nearly sufficient (αs =0.65–0.69).

Keywords: dark triad, PID-5, detachment, deception, confidence judgments, response bias

OPEN ACCESS

Edited by:
John F. Rauthmann,
Wake Forest University, United States

Reviewed by:
Jennifer Vonk,
Oakland University, United States
Konrad Senf,
University of Hohenheim, Germany
Phil Kavanagh,
University of South Australia, Australia

***Correspondence:**
Benno G. Wissing
wissing@uni-kassel.de

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *13 April 2017*
Accepted: *25 August 2017*
Published: *21 September 2017*

Citation:
Wissing BG and Reinhard M-A (2017) The Dark Triad and the PID-5 Maladaptive Personality Traits: Accuracy, Confidence and Response Bias in Judgments of Veracity. Front. Psychol. 8:1549. doi: 10.3389/fpsyg.2017.01549

INTRODUCTION

Research on the Dark Triad (Paulhus and Williams, 2002)—the moderately intercorrelated personality traits of narcissism, Machiavellianism and psychopathy—has accumulated over the recent years. Within the two dominant personality frameworks, the Five-Factor Model (FFM; Costa and McCrea, 1992) and the HEXACO model (Lee and Ashton, 2005), they converge on a core of low agreeableness (Paulhus and Williams, 2002) and low Honesty-Humility (Lee and Ashton, 2005). On the interpersonal level, individuals high in Dark Triad traits are more agentic and lower in communion (Jonason et al., 2010; Jones and Paulhus, 2010) reflecting a low manifestation of agreeableness, which is defined as the willingness to cooperate and, therefore, of central importance for group inclusion (Buss, 1991).

Narcissism is characterized by grandiosity, entitlement, dominance and superiority (Raskin and Hall, 1979; Corry et al., 2008); Machiavellianism is associated with a cold, cynical, amoral worldview and detached, strategic manipulativeness (Christie and Geis, 1970); psychopathy is linked with impulsivity, risk-taking, low neuroticism and low empathy (Hare, 1985). Some authors argue for the domain specific adaptiveness of the Dark Triad traits (e.g., Jonason et al., 2014a).

The existence of different fitness consequences for different personality traits in different environmental niches, suggests that "dark" or "maladaptive" personality traits may represent frequency-dependent fitness optima in certain environmental niches (Penke et al., 2007). Fitness and social desirability are distinct concepts (Nettle, 2006). For example, high Antagonism might be adaptive in exploitable and exploitative environments, exemplified by individuals high in Machiavellianism who disregard conventional morality by rationally defecting when defection is the equilibrium strategy (Gunnthorsdottir et al., 2002).

The integration of individual differences in terms of deception ability within an evolutionary framework varies with its definition of deception ability as a target space for natural selection, as either divisible into deception production and detection ability (Mealey, 1995) or indivisible, conceptualized as a deception-general ability (Wright G. R. T. et al., 2012). In the former case, the dyadic dynamics of a co-evolutionary arms race between predatory, defecting cheaters and cooperators arises (Dawkins and Krebs, 1979; Mealey, 1995)—cheater detection is conceptualized as an evolved mechanism to protect against exploitation in social exchange situations (Cosmides and Tooby, 1992)—, in the latter case "wizards" of deception detection *and* production should result. Empirical support for a deception-general ability currently only exists in the form of found negative correlations between detectability as sender and discrimination ability as receiver (rs $= -0.35, -0.47$; Wright G. R. T. et al., 2012; Wright et al., 2015).

Overall, the data suggest, that humans are only slightly better than chance, at detecting deception (Bond and DePaulo, 2006) and truth-biased in their response, i.e., they believe in the truthfulness of others independently of their actual truthfulness (Levine et al., 1999). On the level of judging deception, humans differ more in response bias than in actual ability (Bond and DePaulo, 2008). Data on the relationship between personality and deception detection accuracy is sparse (Aamodt and Custer, 2006). This sparsity seems to be particularly present concerning the study of dark personality traits and is additionally amplified in terms of the interpretation of results by studies investigating a singular trait without controlling for the shared variance of dark personality traits, e.g., the Dark Triad traits. The results in the existing literature are mixed. For instance, in men, primary psychopathy has been found to be correlated with lie detection ability (Lyons et al., 2013), other studies did not find an association (e.g., Peace and Sinclair, 2012). Also, no superior lie detection ability for Machiavellianism has been found (Zuckerman et al., 1981), but in woman Machiavellianism has been found to be associated with lie detection ability (Lyons et al., 2017). A recent study found no association between Dark Triad traits and deception detection or deception production ability in an interactive deception task (Wright et al., 2015). While the ecological validity of the study (Wright et al., 2015) might be high in comparison with classical studies of deception detection based on audiovisual stimulus material, the statistical power is questionable given the relatively small sample size ($N = 75$) and necessitates further investigation.

Recent research on the Dark Triad traits and active deception indicates that individuals high in Dark Triad traits, particularly those scoring high on the more antagonistic traits of Machiavellianism and psychopathy, differ more from individuals low in Dark Triad traits in deception production frequency (higher for Machiavellianism and psychopathy; Kashy and DePaulo, 1996: Baughman et al., 2014; Jonason et al., 2014b) and amplitude (high-stakes deception for Machiavellianism; Azizli et al., 2016). In contrast, narcissism is primarily associated with self-deception (e.g., Paulhus and Williams, 2002), theorized as an evolutionary evolved intrapersonal mechanism to assist interpersonal deception (von Hippel and Trivers, 2011). The data on the relation between the Dark Triad traits and self-reported lying skills are mixed (Baughman et al., 2014; Jonason et al., 2014b), but suggest overall that the Dark Triad traits are associated with self-perceived deceptive abilities (Giammarco et al., 2013).

Based on findings in the cognitive branches of psychology and neuroscience, the proposed deception-general ability is centered around the associations of executive functions and theory of mind—the ability to understand others' mental states—with deception production and deception detection ability (Wright G. R. T. et al., 2012). Among the Dark Triad traits, grandiose narcissism is exclusively positively associated with theory of mind (Vonk et al., 2015), whereas Machiavellianism and psychopathy are both associated with deficits in empathy and theory of mind (Ali et al., 2009; Ali and Chamorro-Premuzic, 2010; Vonk et al., 2015). Narcissism has also been found to predict self-estimated mind-reading performance (Ames and Kammrath, 2004).

Linked with dark personality traits are maladaptive ones (Grigoras and Wille, 2017). Section III of the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5; American Psychiatric Association, 2013) contains an empirically derived, dimensional model of five maladaptive personality traits (PID-5; Krueger et al., 2012) that constitute the maladaptive versions of the normative Five-Factor Model (FFM; Costa and McCrea, 1992) in the following hierarchical order: Negative Affectivity (i.e., emotional lability, anxiousness, separation insecurity; FFM Neuroticism), Detachment (i.e., withdrawal, anhedonia, intimacy avoidance; low FFM Extraversion), Antagonism (i.e., Manipulativeness, Deceitfulness, Grandiosity; low FFM Agreeableness), Disinhibition (i.e., irresponsibility, impulsivity, distractibility; low FFM Conscientiousness), and Psychoticism (i.e., unusual beliefs and experiences, eccentricity, perceptual dysregulation; FFM Openness; Thomas et al., 2012). Links of the PID-5 traits with narcissism (Wright et al., 2013) and psychopathy (Strickland et al., 2013; Anderson et al., 2014) are established and highlight Antagonism as the central factor. The PID-5 traits have been shown to outperform the Big Five as predictors of the Dark Triad traits (Grigoras and Wille, 2017).

One possible solution to the often-unsuccessful linkage of personality and deception may be to specifically capture the meta-analytically distilled differences on multiple levels (e.g., modality, sender) on the level of receiver-personality. For the most significant individual difference—sender credibility (Bond and DePaulo, 2008)—this has been realized on the level of the receiver by suspiciousness, which has been shown to decrease truth bias if experimentally induced (e.g., McCornack and Levine, 1990; Millar and Millar, 1997; Kim and Levine, 2011). Suspiciousness is one facet of PID-5 Detachment, but the Detachment domain could contain deception relevant features beyond suspiciousness. Meta-analytically extracted modality-based differences in deception detection accuracy and response bias (Bond and DePaulo, 2006) can be interpreted to some extent as the result of modality-mediated differences in detachment between sender and receiver (Burgoon et al., 2008). These differences in sender-receiver-detachment may not be entirely modality-based, but may be partially determined by Detachment on the level of receiver-personality. Therefore, the PID-5 domain of Detachment could play an important role on the level of personality for the process of deception detection.

In this study, the relations between the Dark Triad traits, the PID-5 maladaptive personality traits, and the process of lie detection including detection accuracy, response bias and confidence judgments and process measures for self-reported cue reliance and self-reported decision time were investigated.

Based on the associations of Machiavellianism with a cynical worldview (Christie and Geis, 1970; Jones and Paulhus, 2009), negative views of others (Mudrack, 1993; Rauthmann and Will, 2011; Rauthmann, 2012) and lie acceptability (Wright et al., 2015), Machiavellianism was predicted to be negatively associated with response bias.

Based on its relation with suspiciousness and its potential sender-receiver-detachment enhancing function, Detachment was predicted to be negatively associated with response bias.

Based on the association of the Dark Triad traits with intra- or interpersonal deception production frequency and based on the assumption that frequency is associated with ability, given that deception production ability can be trained (Verschuere et al., 2011; Hu et al., 2012) and, finally, based on the association of deception production ability and deception detection ability (Wright G. R. T. et al., 2012; Wright et al., 2015) it was predicted that the Dark Triad traits are associated with deception detection accuracy.

Based on the negative association of the Dark Triad traits with agreeableness and humility and previous findings regarding self-perceived deceptive competence (Giammarco et al., 2013), it was predicted that the Dark Triad traits are linked to local and global confidence judgments.

Based on the grandiose aspects of Antagonism, it was predicted that Antagonism is associated with local and global confidence judgments. Based on the negative association of Neuroticism with confidence (e.g., Burns et al., 2016), it was predicted that Negative Affectivity is negatively associated with local and global confidence judgments.

MATERIALS AND METHODS

The study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) and the American Psychological Association (APA). Moreover, by the time the data were acquired it was also not required at Kassel University, nor at most other German universities to seek ethics approval for simple studies on personality and attitudes. The study exclusively makes use of anonymous questionnaires. No identifying information was obtained from participants. The participants were explicitly informed that the data are treated confidentially. Every participant had to agree to the following statements: "I understand that my participation is voluntary and that I may withdraw from the study at any time without explanation;" and "I hereby confirm that I am at least 18 years old, and that I agree to take part in this study." Furthermore, they could withdraw from the study at any time.

Statistical Power and Participants

Based on effects sizes from previous studies for deception detection accuracy (Bond and DePaulo, 2006) and response bias (Bond and DePaulo, 2008), we estimated a lower sample size bound of $N = 176/204$ based on a small to medium effect size of $f^2 = 0.1$ with $k = 3/5$ predictors, $\alpha = 0.05$, Power $1–\beta = 0.95$ using the statistical power analysis tool G*Power (Faul et al., 2009). Given only the small meta-analytically identified interindividual differences in accuracy (Bond and DePaulo, 2006), this effect size may still be appointed too high for accuracy, but considering the proposed general-deception ability (Wright G. R. T. et al., 2012) it seems more reasonable in the context of traits associated with high intra- or interpersonal deception frequency—given the assumption that frequency is associated with ability, given that deception production ability can be trained (Verschuere et al., 2011; Hu et al., 2012)—and the estimated sample size is a significant improvement in terms of statistical power over the original study ($N = 75$; Wright et al., 2015).

Participants that dropped out before finishing the deception detection task were excluded from data analysis, resulting in the final sample of 207 participants (59.9% female; *M* age = 29.03; *SD* age = 10.62, age range = 17–66) that were recruited from Germany using online invocations and invitations on the campus of the university of Kassel. The study was conducted online and lasted for a total of approximately 25 min.

Procedure and Measures

Personality was assessed using the German version of the *Naughty Nine* short scale, a 9-item psychometrically optimized version of the *Dirty Dozen* (Jonason and Webster, 2010) self-report instrument, which measures the Dark Triad traits with good internal consistency and stability (Küfner et al., 2015), consisting of narcissism (e.g., "*I tend to want others to admire me*"; $\alpha = 0.82$), Machiavellianism (e.g., "*I tend to manipulate others to get my way*"; $\alpha = 0.75$) and psychopathy (e.g., "*I tend to lack remorse*"; $\alpha = 0.69$) using a 9-point assumed interval-type scale (1 = *disagree strongly*, 9 = *agree strongly*), followed by the German version of the *Personality Inventory for DSM-5 Brief Form* (*PID-5-BF*; Krueger et al., 2012; American Psychiatric Association, 2013), that measures the five maladaptive dimensional personality trait domains of the PID-5 model with 25 items, consisting of Negative Affectivity (e.g., "*I worry about almost everything*"; $\alpha = 0.68$), Detachment (e.g., "*I often feel like nothing I do really matters*"; $\alpha = 0.65$), Antagonism (e.g., "*It's no big deal if I hurt other peoples' feelings*"; α = 0.72), Disinhibition (e.g., "*People would describe me as reckless*"; α = 0.69), and Psychoticism (e.g., "*My thoughts often don't make sense to others*"; α = 0.77) using a 4-point Likert-type scale (0 = *very false/often very false*, 3 = *very true/often true*). The US and Danish version of the PID-5-BF have shown acceptable internal consistencies (Anderson et al., 2016; Bach et al., 2016). The full 220-item version of the PID-5 is currently validated in German (Zimmermann et al., 2014).

To measure deception detection accuracy, participants were randomly assigned to one of two video sets. Each video set consisted of 14 videos, in one half of which the sender was telling the truth, whereas telling lies in the other half, so that the information that was depicted (honestly vs. dishonestly) was balanced across the videos assigned to different participants. 4 out of 14 senders of each video set were female. The videos displayed an employment interview situation, where the candidate (sender) was asked a question by the interviewer (receiver), whereas only the candidate was visible and the camera perspective simulated the point of view of the interviewer. The entire body of the sender was visible and the interviewer was blind to the experimental conditions. Each sender was instructed to convince the interviewer of a job they had in the past vs. one they had not worked in in the past. Every sender was only visible in one of the two video sets (further information on the audiovisual stimulus material can be found in Reinhard et al., 2013). To measure deception detection accuracy, the participants were instructed to decide, whether the candidate was telling the truth or lying after watching each video. After each binary truth vs. lie decision was made, the participants were asked how confident they were in their judgment (e.g., "*How confident are you in your judgment?*"; α = 0.84) using a continuous percental-type scale (0% = *absolutely uncertain*, 100% = *absolutely certain*). This process was repeated for the totality of all 14 videos. Truth bias was used as the response bias measure and was determined by the total number of truth judgments.

After completion of the deception detection task, self-reported decision time was measured (e.g., "*When did you decide on the truthfulness of the candidate?*") by using an 11-point assumed interval-type scale (0 = *directly at the beginning of the video*, 10 = *after completion of the video*). Thereafter, self-reported verbal and nonverbal cue reliance were measured on a 10-point assumed interval-type scale (1 = *strongly disagree*, 10 = *strongly agree*) with two items each (e.g., "*I focused on the content*," "*I used the content of the statements for my judgment*"; α = 0.94; "*I focused on the nonverbal behavior*, "*I used the nonverbal behavior for my judgment*"; α = 0.90). Finally, to measure global confidence, the participants were asked to estimate their overall detection accuracy in absolute terms (e.g., "*How many of the 14 videos do you think you judged correctly?*"; 0–14).

RESULTS

Because of multiple comparisons, a significance threshold of α = 0.01 was used to reduce Type I errors. All statistical statements are relating to the data, not the theory. Data analysis was conducted with R (R Core Team, 2017).

Prespecified Data Analysis

Response Bias

Zero-order correlations and standardized regression weights for the personality variables and response bias can be seen in **Tables 1**, **2**. The participants were truth-biased [M = 64.63%, SD = 17.45; $M\Delta$ = 14.63, 95% CI = (12.24, 17.02); $t_{(206)}$ = 12.06, p < 0.001]. The data did not support the predicted negative association of Machiavellianism with response bias [$r_{(205)}$ = −0.05, 95% CI = (−0.19, 0.08), p = 0.448]; However, the predicted negative association between Detachment and response bias was supported by the data [$r_{(205)}$ = −0.23, 95% CI = (−0.36, −0.10), p < 0.001].

A regression model with Detachment as the predictor variable and response bias as criterion variable, suggested substantial heteroscedasticity of the residuals. Therefore, a robust regression with MM-estimator was computed. As seen in **Figure 1**, Detachment emerged as a substantial predictor [$b = -7.57$, 95% $CI = (-11.42, -3.71)$] that predicted a response bias range of $\hat{y} = [49.54, 72.25]$; With all PID-5 traits entered as predictor variables into the model, Detachment emerged as the only substantial predictor [$b = -7.50$, 95% $CI = (-11.97, -3.04)$] and showed no substantial difference in its association pattern with the criterion variable.

TABLE 1 | Zero-order correlations and standardized regression weights with 95% CIs (in brackets) for the Dark Triad and deception variables.

	R^2 (f^2)	Narcissism	Machiavellianism	Psychopathy	Dark Triad composite
DECEPTION DETECTION ACCURACY					
Overall	0.00 (−0.01; 0.00; 0.02)	−0.03 (−0.19; −0.03; 0.13)	−0.02 (−0.16; 0.01; 0.17)	−0.05 (−0.19; −0.05; 0.10)	−0.05
Truth	0.02 (−0.02; 0.02; 0.07)	0.00 (−0.13; 0.03; 0.18)	−0.06 (−0.18; −0.02; 0.14)	−0.15 (−0.30; −0.15; −0.01)	−0.09
Lie	0.01 (−0.02; 0.01; 0.05)	−0.03 (−0.22; −0.06; 0.10)	0.03 (−0.13; 0.03; 0.19)	0.11 (−0.04; 0.11; 0.25)	0.05
Response bias	0.03 (−0.02; 0.03; 0.07)	0.02 (−0.11; 0.05; 0.21)	−0.05 (−0.19; −0.03; 0.13)	−0.15 (−0.30; −0.15; −0.01)	−0.08
CONFIDENCE JUDGMENTS					
Local	0.02 (−0.02; 0.03; 0.07)	0.06 (−0.14; 0.02; 0.17)	0.11 (−0.09; 0.07; 0.23)	0.14 (−0.03; 0.11; 0.25)	0.14
Global	0.05 (0.00; 0.06; 0.13)	0.12 (−0.10; 0.05; 0.21)	0.16 (−0.08; 0.09; 0.25)	0.20* (0.03; 0.17; 0.31)	0.22*
Over	0.05 (−0.01; 0.05; 0.12)	0.12 (−0.09; 0.06; 0.21)	0.15 (−0.09; 0.07; 0.23)	0.20* (0.03; 0.17; 0.31)	0.21*

*N = 207; *p < 0.01 (two-tailed).*

TABLE 2 | Zero-order correlations and standardized regression weights with 95% CIs (in brackets) for the PID-5 traits and deception variables.

	***R*² (*f*²)**	**Negative Affectivity**	**Detachment**	**Antagonism**	**Disinhibition**	**Psychoticism**
DECEPTION DETECTION ACCURACY						
Overall	0.01 (−0.02; 0.01; 0.05)	−0.05 (−0.22; −0.06; 0.10)	−0.02 (−0.15; 0.01; 0.17)	−0.09 (−0.25; −0.10; 0.05)	0.01 (−0.10; 0.05; 0.21)	−0.04 (−0.18; −0.01; 0.17)
Truth	0.06 (0.00; 0.07; 0.14)	0.04 (0.00; 0.16; 0.31)	−0.21* (−0.35; −0.19; −0.04)	−0.08 (−0.16; −0.01; 0.13)	−0.07 (−0.19; −0.04; 0.11)	−0.13 (−0.27; −0.10; 0.08)
Lie	0.08* (0.01; 0.09; 0.18)	−0.10 (−0.38; −0.23*; −0.07)	0.19* (0.06; 0.21*; 0.37)	−0.02 (−0.24; −0.10; 0.05)	0.09 (−0.05; 0.10; 0.24)	0.09 (−0.08; 0.09; 0.26)
Response bias	0.09* (0.02; 0.10; 0.20)	0.08 (0.07; 0.23*; 0.38)	−0.23** (−0.39; −0.24*; −0.09)	−0.04 (−0.09; 0.05; 0.19)	−0.10 (−0.23; −0.08; 0.07)	−0.13 (−0.28; −0.11; 0.06)
CONFIDENCE JUDGMENTS						
local	0.03 (−0.01; 0.03; 0.08)	−0.14 (−0.29; −0.13; 0.03)	−0.07 (−0.21; −0.05; 0.10)	0.07 (−0.05; 0.10; 0.24)	−0.04 (−0.15; 0.00; 0.15)	−0.06 (−0.18; 0.00; 0.17)
global	0.05 (0.00; 0.06; 0.13)	−0.12 (−0.22; −0.07; 0.09)	−0.12 (−0.26; −0.10; 0.05)	0.12 (0.03; 0.17; 0.32)	−0.02 (−0.12; 0.04; 0.19)	−0.13 (−0.28; −0.11; 0.06)
over	0.05 (−0.01; 0.05; 0.12)	−0.08 (−0.18; −0.02; 0.13)	−0.09 (−0.25; −0.09; 0.06)	0.15 (0.06; 0.20*; 0.35)	−0.02 (−0.15; 0.00; 0.15)	−0.09 (−0.26; −0.09; 0.08)

*Note: N = 207; *p < 0.01 **p < 0.001 (two-tailed).*

Linear model assumptions were validated including by using the R package gvlma (Pena and Slate, 2014). The robust regression model was computed by using the R packages MASS (Venables and Ripley, 2002), robust (Wang et al., 2017), and prediction (Leeper, 2017). **Figure 1** was constructed with the R package visreg (Breheny and Burchett, 2017) and the ggplot2 (Wickham, 2009) plotting engine.

Deception Detection Accuracy

Overall detection accuracy was $M = 50.38\%$ ($SD = 11.00$) and not different from chance [$t_{(206)} = 0.50$, $p = 0.620$]. Zero-order correlations and standardized regression weights for the personality variables with measures of deception detection accuracy can be seen in **Tables 1**, **2**. The data did not support the predicted association between the Dark Triad traits and deception detection accuracy.

Confidence Judgments

Zero-order correlations and standardized regression weights for the personality variables with measures of confidences judgments can be seen in **Tables 1**, **2**. Among the Dark Triad traits, a substantial association with confidence judgments emerged only for psychopathy and global confidence judgments [$r_{(205)} = 0.20$, 95% $CI = (0.07, 0.33)$, $p = 0.004$]. Antagonism and Neuroticism were not substantially associated with confidence judgments.

Exploratory Data Analysis

Personality

Intercorrelations of the personality traits can be seen in **Table 3**. As expected based on prior research, all individual Dark Triad traits were associated most strongly with Antagonism. Overall, the correlational pattern is in line with the one found in a previous study with longer measures of the Dark Triad traits and the PID-5 traits (Grigoras and Wille, 2017). Personality scales were computed with the R package psych (Revelle, 2017).

Truth and Lie Detection Accuracy

Zero-order correlations and standardized regression weights for the personality variables and truth and lie detection accuracy can be seen in **Tables 1**, **2**. Accuracy for truth detection was above chance [$M = 65.01\%$, $SD = 21.05$; $M\Delta = 15.01$, 95% $CI = (12.13, 17.89)$; $t_{(206)} = 10.26$, $p < 0.001$]. Detachment was negatively associated with truth detection accuracy [$r_{(205)} = -0.21$, 95% $CI = (-0.33, -0.07)$, $p = 0.003$]. When controlling for response bias in a first-order partial correlation, the confidence interval for the associations between Detachment and truth detection accuracy included zero [$r_{(204)} = -0.01$, 95% $CI = (-0.15, 0.12)$, $p = 0.850$]. Accuracy for lie detection was below chance [$M = 35.75\%$, $SD = 20.20$; $M\Delta = -14.25$, 95% $CI = (-17.02, -11.48)$; $t_{(206)} = -10.15$, $p < 0.001$]. Detachment was positively associated with lie detection accuracy [$r_{(205)} = 0.19$, 95% $CI = (0.05, 0.32)$, $p = 0.006$]. When controlling for response bias in a first-order partial correlation, the confidence interval for the associations between Detachment and lie detection accuracy included zero [$r_{(204)} = -0.01$, 95% $CI = (-0.15, 0.12)$, $p = 0.850$]. First-order partial correlations were computed with the R package psych (Revelle, 2017).

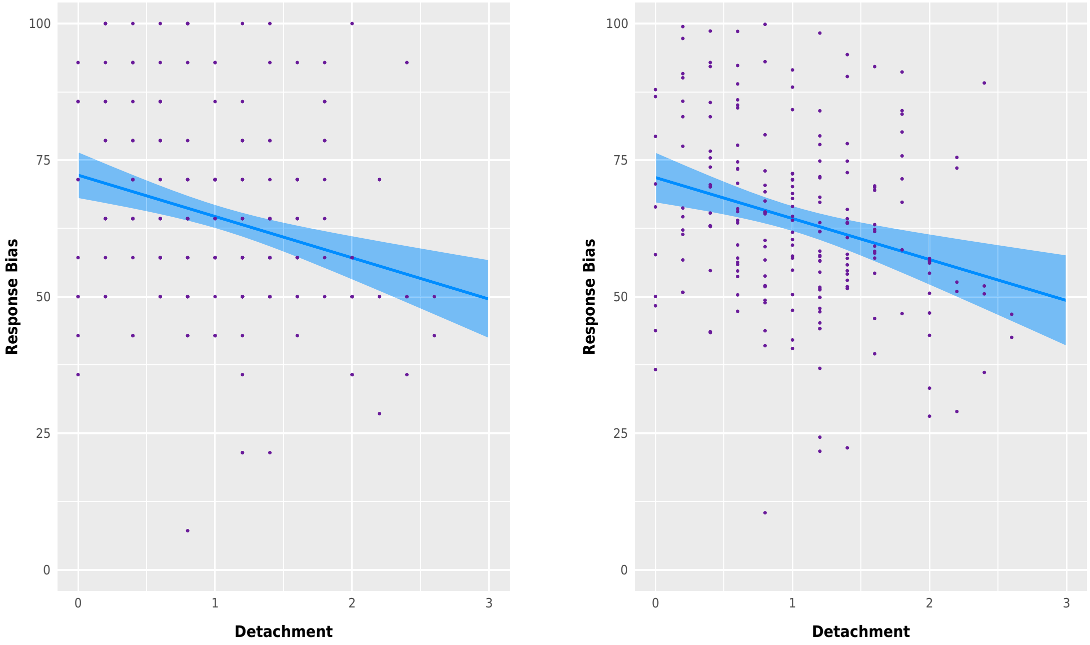

FIGURE 1 | Robust linear regression with 95% CI for response bias as a function of Detachment with $k = 1$ predictor **(left)** and all $k = 5$ dimensions of PID-5 maladaptive personality space as predictors **(right)**.

TABLE 3 | Descriptive statistics and zero-order correlations for the Dark Triad traits and the PID-5 traits.

	α	*M* (*SD*)	1	2	3	4	5	6	7	8	9
DARK TRIAD											
1. Narcissism	0.82	4.83 (1.92)	–								
2. Machiavellianism	0.75	3.97 (1.90)	0.48**	–							
3. Psychopathy	0.69	3.17 (1.77)	0.13	0.32**	–						
4. Dark Triad composite	0.78	3.99 (1.37)	0.75**	0.82**	0.64**	–					
PID-5											
5. Negative Affectivity	0.68	1.28 (0.59)	0.15	0.09	−0.21*	0.02	–				
6. Detachment	0.65	1.04 (0.61)	−0.09	0.10	0.22*	0.10	0.30**	–			
7. Antagonism	0.72	0.60 (0.50)	0.35**	0.62**	0.54**	0.68**	0.08	0.25**	–		
8. Disinhibition	0.69	0.83 (0.55)	0.00	0.14	0.16	0.13	0.35**	0.27**	0.21*	–	
9. Psychoticism	0.77	1.02 (0.67)	0.11	0.26**	0.06	0.20*	0.47**	0.46**	0.28**	0.38**	–

N = 207; *p < 0.01 **p < 0.001 (two-tailed).

Confidence Judgments

To further explore confidence judgments and deception detection performance, a measure of overconfidence was computed by subtracting global confidence judgments (the number of self-estimated correct judgments) by the number of actual accurate judgments. As can be seen in **Tables 1**, **2**, psychopathy and the Dark Triad composite were associated with overconfidence. Antagonism emerged as a predictor of overconfidence.

Self-reported Process Measures

Personality variables were not substantially associated with self-reported process measures. In line with previous findings (Reinhard et al., 2011), self-reported verbal cue reliance was associated with overall deception detection accuracy [$r_{(205)} = 0.22$, 95% $CI = (0.08, 0.34)$, $p = 0.002$].

DISCUSSION

The present study investigated the relation between the Dark Triad traits, the PID-5 maladaptive personality traits and the process of lie detection including detection accuracy, response bias, confidence judgments and process measures for self-reported cue reliance and self-reported decision time.

There was no association of Dark Triad traits with the ability of deception detection in the data. This finding is in line with previous research (e.g., Wright et al., 2015), that found no relation of the Dark Triad traits with deception detection ability. Instead, and also in line with previous findings (Giammarco et al., 2013), an association of psychopathy and the Dark Triad composite with global confidence judgments appeared in the data. More importantly, the global confidence judgments were not grounded in actual deception detection performance: On average, individuals with higher psychopathy and overall Dark Triad scores reported higher confidence in their deception detection accuracy than their actual accuracy permitted—they were overconfident in their ability. The confidence interval of the standardized regression coefficients for psychopathy suggested, that psychopathy alone could account for unique variance in global confidence and overconfidence above the shared variance of the Dark Triad. This pattern is in line with findings of a recent meta-analysis, that psychopathy is often the only significant correlate of important psychosocial outcomes, if the shared variance of the Dark Triad is controlled (Muris et al., 2017). The absence of a substantial association pattern between Machiavellianism and response bias should be interpreted cautiously, given the relatively high prior probability in the form of the well-established connection between Machiavellianism and interpersonal suspiciousness.

On the level of maladaptive personality, Detachment emerged as a predictor of response bias with a minus-signed coefficient. Effect sizes for response bias predicted by Detachment can be considered substantial in the context of the criterion construct, given that humans are generally truth-biased and that very high Detachment scores predicted the absence of response bias. In a meta-analysis with 32 samples, a mean observed standard range of 50.06% in response bias was found (Bond and DePaulo, 2008). The response bias range predicted by Detachment was 22.71%, which corresponds to 45.37% of the meta-analytically mean observed standard range in response bias.

While the truth-default mode is likely to be adaptive in environments, where most communication is honest (Levine, 2014), Detachment may facilitate adaptive behavior in environments with high deception frequency by providing a lower or even no response bias, and therefore, a higher lie detection accuracy. In such environments, the interpersonally active core of PID-5 Detachment—withdrawal from other people—may not be maladaptive, but serve a protective function. This potential adaptive function in environmental niches with high deception frequency is contrasted by the finding that within PID-5 maladaptive personality space, facets of Detachment and Negative Affectivity exhibit the strongest connections with a general index of personality disorder severity (Hopwood et al., 2012).

Limitations and Future Research

The present study was based on short self-report measures. The self-reported measures of cue reliance and decision time are inherently subjective. It is highly questionable, if the subjects had cognitive access to their cue reliance modalities, but self-reported cue reliance is associated with objective outcomes, e.g., verbal cue reliance is associated with deception detection accuracy (Reinhard et al., 2011). On the level of personality assessment, the short Naughty Nine instrument revealed an anomalistic intercorrelation pattern of the Dark Triad traits, in which Machiavellianism was more strongly associated with narcissism than with psychopathy. The PID-5-BF instrument can only measure maladaptive personality on the level of domains and can potentially produce a higher measurement error in subclinical samples (Krueger and Markon, 2014). Psychopathy, Negative Affectivity, Detachment, and Disinhibition had questionable internal consistencies (αs = 0.65–0.69). The internal consistency coefficients are in line with the corresponding validation studies, e.g., αs = 0.57–0.76 for psychopathy (Küfner et al., 2015) and the validation study of the US version of the PID-5-BF found two alpha coefficients below α = 0.70 (Anderson et al., 2016). Overall, the intended optimization of the trade-off between reliability and the prevention of fatigue effects in the process of deception detection was not sufficiently successful. Beyond these limitations, the generalizability of the findings to real-world contexts of deception detection is questionable. Furthermore, the non-significant results could represent type II errors resulting from power deficiency, given that a true effect exits with a smaller size than estimated. Future studies should therefore aim to use longer instruments with higher internal consistencies and larger samples to increase statistical power.

In the current PID-5 model, suspiciousness is a facet of both Detachment and Negative Affectivity, but the replicated finding of marginal secondary loadings of suspiciousness on Negative Affectivity questions the relationship of the facet with its higher-order factor (Wright A. G. C. et al., 2012; De Fruyt et al., 2013; Zimmermann et al., 2014). An interesting question is, if facets of Detachment beyond suspiciousness can account for variance in response bias. Future research should use longer versions of the PID-5, which can measure the facets of Detachment. Beyond the facet level resolution, the interactions between modality-based and personality-based differences in Detachment are worth investigating. Are there levels of Detachment on the level of personality that can account for modality-based differences in sender-receiver-detachment for response bias and deception detection accuracy? In situations where cues are impairing deception detection accuracy, individuals high in Detachment may find it easier to ignore these unreliable cues to deception.

The relation of the Dark Triad traits and deception ability could be investigated in experimental settings, that provide a more optimal fit regarding Dark Triad specific motivations and affordances, e.g., given that situational familiarity enhances deception detection accuracy (Reinhard et al., 2011). The Dark Triad traits should express their antagonistic behaviors specifically in selfishness vs. cooperation scenarios (Rauthmann, 2012), driven by the shared psychogenic motivational core of power (Kajonius et al., 2015; Jonason and Ferrell, 2016). Future studies should, therefore, strive to activate the power motive by providing incentives for power acquisition via deception production or deception detection in contexts, that provide selfish vs. cooperative behavioral optionality.

STATEMENT FOR DISCLOSURE OF SAMPLE, CONDITIONS, MEASURES, AND EXCLUSIONS

The authors confirm that they have reported all measures, conditions, data exclusions, and how they determined their sample size.

CODE, DATA AND MATERIALS

Code, data and materials can be accessed via: https://osf.io/tcy3q/.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

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

Aamodt, M. G., and Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner* 15, 6–11.

Ali, F., Amorim, I. S., and Chamorro-Premuzic, T. (2009). Empathy deficits and trait emotional intelligence in psychopathy and Machiavellianism. *Pers. Indiv. Differ*. 47, 758–762. doi: 10.1016/j.paid.2009.06.016

Ali, F., and Chamorro-Premuzic, T. (2010). Investigating theory of mind deficits in nonclinical psychopathy and Machiavellianism. *Pers. Indiv. Differ*. 49, 169–174. doi: 10.1016/j.paid.2010.03.027

Ames, D. R., and Kammrath, L. K. (2004). Mind-reading and metacognition: narcissism, not actual competence, predicts self-estimated ability. *J. Nonverbal Behav*. 28, 187–209. doi: 10.1023/B:JONB.0000039649.20015.0e

Anderson, J. L., Sellbom, M., and Salekin, R. T. (2016). Utility of the Personality Inventory for DSM-5-Brief Form (PID-5-BF) in the measurement of maladaptive personality and psychopathology. *Assessment* doi: 10.1177/1073191116676889. [Epub ahead of print].

Anderson, J. L., Sellbom, M., Wygant, D. B., Salekin, R. T., and Krueger, R. F. (2014). Examining the associations between DSM-5 section III antisocial personality disorder traits and psychopathy in community and university samples. *J. Pers. Disord.* 28, 675–697. doi: 10.1521/pedi_2014_28_134

Azizli, N., Atkinson, B. E., Baughman, H. M., Chin, K., Vernon, P. A., Harris, E., et al. (2016). Lies and crimes: dark triad, misconduct, and high-stakes deception. *Pers. Indiv. Differ*. 89, 34–39. doi: 10.1016/j.paid.2015.09.034

Bach, B., Maples-Keller, J. L., Bo, S., and Simonsen, E. (2016). The alternative DSM-5 personality disorder traits criterion: a comparative examination of three self-report forms in a Danish population. *Personal. Disord.* 7, 124–135. doi: 10.1037/per0000162

Baughman, H. M., Jonason, P. K., Lyons, M., and Vernon, P. A. (2014). Liar liar pants on fire: cheater strategies linked to the dark triad. *Pers. Indiv. Differ*. 71, 35–38. doi: 10.1016/j.paid.2014.07.019

Bond, C. F., and DePaulo, B. M. (2006). Accuracy of deception judgments. *Pers. Soc. Psychol. Rev*. 10, 214–234. doi: 10.1207/s15327957pspr1003_2

Bond, C. F., and DePaulo, B. M. (2008). Individual differences in judging deception: accuracy and bias. *Psychol. Bull.* 134, 477–492. doi: 10.1037/0033-2909.134.4.477

Breheny, P., and Burchett, W. (2017). *visreg: Visualization of Regression Models.* R package version 2.4–1. Available online at: https://CRAN.R-project.org/package=visreg

Burgoon, J. K., Blair, J. P., and Strom, R. E. (2008). Cognitive biases and nonverbal cue availability in detecting deception. *Hum. Commun. Res.* 34, 572–599. doi: 10.1111/j.1468-2958.2008.00333.x

Burns, K. M., Burns, N. R., and Ward, L. (2016). Confidence –more a personality or ability trait? it depends on how it is measured: a comparison of young and older adults. *Front. Psychol.* 7:518. doi: 10.3389/fpsyg.2016.00518

Buss, D. M. (1991). Evolutionary personality psychology. *Annu. Rev. Psychol.* 42, 459–491. doi: 10.1146/annurev.ps.42.020191.002331

Christie, R., and Geis, F. L. (1970). *Studies in Machiavellianism*. New York, NY: Academic Press.

Corry, N., Merritt, R. D., Mrug, S., and Pamp, B. (2008). The factor structure of the narcissistic personality inventory. *J. Pers. Assess.* 90, 593–600. doi: 10.1080/00223890802388590

Cosmides, L., and Tooby, J. (1992). Cognitive adaptations for social exchange. *Adapted Mind* 31, 163–228.

Costa, P. T., and McCrea, R. R. (1992). *NEO Five-Factor Inventory*. Lutz, FL: Psychological Assessment Resources.

Dawkins, R., and Krebs, J. R. (1979). Arms races between and within species. *Proc. R. Soc. B Biol. Sci.* 205, 489–511. doi: 10.1098/rspb.1979.0081

De Fruyt, F., De Clercq, B., De Bolle, M., Wille, B., Markon, K., and Krueger, R. F. (2013). General and maladaptive traits in a fivefactor framework for DSM-5 in a university student sample. *Assessment* 20, 295–307. doi: 10.1177/1073191113475808

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Giammarco, E. A., Atkinson, B., Baughman, H., Veselka, L., and Vernon, P. A. (2013). The relation between antisocial personality and the perceived ability to deceive. *Pers. Indiv. Differ*. 54, 246–250. doi: 10.1016/j.paid.2012.09.004

Grigoras, M., and Wille, B. (2017). Shedding light on the dark side: associations between the dark triad and the DSM-5 maladaptive trait model. *Pers. Indiv. Differ*. 104, 516–521. doi: 10.1016/j.paid.2016.09.016

Gunnthorsdottir, A., McCabe, K., and Smith, V. (2002). Using the Machiavellianism instrument to predict trustworthiness in a bargaining game. *J. Econ. Psychol.* 23, 49–66. doi: 10.1016/S0167-4870(01)00067-8

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *J. Consult. Clin. Psych.* 53, 7–16. doi: 10.1037/0022-006X.53.1.7

Hopwood, C. J., Thomas, K. M., Markon, K. E., Wright, A. G. C., and Krueger, R. F. (2012). DSM-5 personality traits and DSM-IV personality disorders. *J. Abnorm. Psychol.* 121, 424– 432. doi: 10.1037/a0026656

Hu, X., Chen, H., and Fu, G. (2012). A repeated lie becomes a truth? The effect of intentional control and training on deception. *Front. Psychol.* 3:488. doi: 10.3389/fpsyg.2012.00488

Jonason, P. K., and Ferrell, J. D. (2016). Looking under the hood: the psychogenic motivational foundations of the dark triad. *Pers. Indiv. Differ*. 94, 324–331. doi: 10.1016/j.paid.2016.01.039

Jonason, P. K., Li, N. P., and Teicher, E. A. (2010). Who is James Bond? The dark triad as an agentic social style. *Individ. Differ. Res.* 8, 111–120.

Jonason, P. K., Wee, S., and Li, N. P. (2014a). Thinking bigger and better about "bad apples": evolutionary industrial/organizational psychology and the dark triad. *Ind. Organ. Psychol.* 7, 117–121. doi: 10.1111/iops.12118

Jonason, P. K., Lyons, M., Baughman, H. M., and Vernon, P. A. (2014b). What a tangled web we weave: the dark triad and deception. *Pers. Indiv. Differ*. 70, 117–119. doi: 10.1016/j.paid.2014.06.038

Jonason, P. K., and Webster, G. D. (2010). The dirty dozen: a concise measure of the dark triad. *Psychol. Assess.* 22, 420–432. doi: 10.1037/a0019265

Jones, D. N., and Paulhus, D. L. (2009). "Machiavellianism," in *Handbook of Individual Differences in Social Behavior*, eds M. R. Leary and R. H. Hoyle (New York, NY: Guilford), 93–108.

Jones, D. N., and Paulhus, D. L. (2010). "Differentiating the Dark Triad within the interpersonal circumplex," in *Handbook of Interpersonal Theory and Research*, eds L. M. Horowitz, and S. N. Strack (New York, NY: Guilford), 249–267.

Kajonius, P. J., Persson, B. N., and Jonason, P. K. (2015). Hedonism, achievement, and power: universal values that characterize the dark triad. *Pers. Indiv. Differ*. 77, 173–178. doi: 10.1016/j.paid.2014.12.055

Kashy, D. A., and DePaulo, B. M. (1996). Who lies? *J. Pers. Soc. Psychol.* 70, 1037–1051. doi: 10.1037/0022-3514.70.5.1037

Kim, R. K., and Levine, T. R. (2011). The effect of suspicion on deception detection accuracy: optimal level or opposing effects. *Commun. Rep.* 24, 51–62. doi: 10.1080/08934215.2011.615272

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., and Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychol. Med.* 42, 1879–1890. doi: 10.1017/S0033291711002674

Krueger, R. F., and Markon, K. E. (2014). The role of the DSM-5 personality trait model in moving toward a quantitative and empirically based approach to classifying personality and psychopathology. *Annu. Rev. Clin. Psychol.* 10, 477–501. doi: 10.1146/annurev-clinpsy-032813-153732

Küfner, A. C. P., Dufner, M., and Back, M. D. (2015). The dirty dozen and the naughty nine–short scales for the assessment of narcissism, Machiavellianism and psychopathy. *Diagnostica* 61, 76–91. doi: 10.1026/0012-1924/a000124

Lee, K., and Ashton, M. C. (2005). Psychopathy, Machiavellianism, and narcissism in the five-factor model and the HEXACO model of personality structure. *Pers. Indiv. Differ.* 38, 1571–1582. doi: 10.1016/j.paid.2004.09.016

Leeper, T. J. (2017). *prediction: Tidy, Type-Safe 'Prediction()' Methods*. R package version 0.2.0.

Levine, T. R. (2014). Truth-Default Theory (TDT): a theory of human deception and deception detection. *J. Lang. Soc. Psychol.* 33, 378–392. doi: 10.1177/0261927X14535916

Levine, T. R., Park, H. S., and McCornack, S. A. (1999). Accuracy in detecting truths and lies: documenting the "veracity effect". *Commun. Monogr.* 66, 125–144. doi: 10.1080/03637759909376468

Lyons, M., Croft, A., Fairhurst, S., Varley, K., and Wilson, C. (2017). Seeing through crocodile tears? Sex-specific associations between the Dark Triad traits and lie detection accuracy. *Pers. Indiv. Differ.* 113, 1–4. doi: 10.1016/j.paid.2017.03.008

Lyons, M., Healy, N., and Bruno, D. (2013). It takes one to know one–psychopathy and deception detection ability. *Pers. Indiv. Differ.* 55, 676–679. doi: 10.1016/j.paid.2013.05.018

McCornack, S. A., and Levine, T. R. (1990). When lovers become leery: the relationship between suspicion and accuracy in detecting deception. *Commun. Monogr.* 57, 219–230. doi: 10.1080/03637759009376197

Mealey, L. (1995). The sociobiology of sociopathy: an integrated evolutionary model. *Behav. Brain Sci.* 18, 523–599. doi: 10.1017/S0140525X00039595

Millar, M. G., and Millar, K. U. (1997). The effects of cognitive capacity and suspicion on response bias. *Commun. Res.* 24, 556–570. doi: 10.1177/009365097024005005

Mudrack, P. E. (1993). An investigation into the acceptability of workplace behaviors of a dubious ethical nature. *J. Bus. Ethics* 12, 517–524. doi: 10.1007/BF00872373

Muris, P., Merckelbach, H., Otgaar, H., and Meijer, E. (2017). The malevolent side of human nature: a meta-analysis and critical review of the literature on the dark triad (Narcissism, Machiavellianism, and Psychopathy). *Perspect. Psychol. Sci.* 12, 183–204. doi: 10.1177/1745691616666070

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *Am. Psychol.* 61, 622–631. doi: 10.1037/0003-066X.61.6.622

Paulhus, D. L., and Williams, K. M. (2002). The Dark triad of personality: narcissism, Machiavellianism, and psychopathy. *J. Res. Pers.* 36, 556–563. doi: 10.1016/S0092-6566(02)00505-6

Peace, K. A., and Sinclair, S. M. (2012). Cold-blooded lie catchers? An investigation of psychopathy, emotional processing, and deception detection. *Legal Criminol. Psych.* 17, 177–191. doi: 10.1348/135532510X524789

Pena, E. A., and Slate, E. H. (2014). *gvlma: Global Validation of Linear Models Assumptions*. R package version 1.0.0.2. Available online at: https://CRAN.R-project.org/package=gvlma

Penke, L., Denissen, J. J. A., and Miller, G. F. (2007). The evolutionary genetics of personality. *Eur. J. Pers.* 21, 549–587. doi: 10.1002/per.629

Raskin, R., and Hall, C. S. (1979). A Narcissistic Personality Inventory. *Psychol. Rep.* 45:590. doi: 10.2466/pr0.1979.45.2.590

Rauthmann, J. F. (2012). The Dark Triad and interpersonal perception: similarities and differences in the social consequences of narcissism, Machiavellianism, and psychopathy. *Soc. Psychol. Pers. Sci.* 3, 487–496. doi: 10.1177/1948550611427608

Rauthmann, J. F., and Will, T. (2011). Proposing a multidimensional Machiavellianism conception. *Soc. Behav. Personal.* 39, 391–404. doi: 10.2224/sbp.2011.39.3.391

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Reinhard, M.-A., Sporer, S. L., Scharmach, M., and Marksteiner, T. (2011). Listening, not watching: situational familiarity and the ability to detect deception. *J. Pers. Soc. Psychol.* 101, 467–484. doi: 10.1037/a0023726

Reinhard, M., Scharmach, M., and Müller, P. (2013). It's not what you are, it's what you know: experience, beliefs, and the detection of deception in employment interviews. *J. Appl. Soc. Psychol.* 43, 467–479. doi: 10.1111/j.1559-1816.2013.01011.x

Revelle, W. (2017). *psych: Procedures for Personality and Psychological Research*. R package version 1.7.5. Evanston, IL: Northwestern University. Available online at: https://CRAN.R-project.org/package=psych

Strickland, C. M., Drislane, L. E., Lucy, M., Krueger, R. F., and Patrick, C. J. (2013). Characterizing psychopathy using DSM-5 personality traits. *Assessment* 20, 327–338. doi: 10.1177/1073191113486691

Thomas, K. M., Yalch, M. M., Krueger, R. F., Wright, A. G. C., Markon, K. E., and Hopwood, C. J. (2012). The convergent structure of DSM-5 personality trait facets and Five-Factor Model trait domains. *Assessment* 20, 308–311. doi: 10.1177/1073191112457589

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. 4th Edn. New York, NY: Springer. ISBN 0-387-95457-0.

Verschuere, B., Spruyt, A., Meijer, E. H., and Otgaar, H. (2011). The ease of lying. *Conscious. Cogn.* 20, 908–911. doi: 10.1016/j.concog.2010.10.023

von Hippel, W., and Trivers, R. (2011). The evolution and psychology of self-deception. *Behav. Brain Sci.* 34, 1–16. doi: 10.1017/S0140525X10001354

Vonk, J., Zeigler-Hill, V., Ewing, D., Mercer, S., and Noser, A. E. (2015). Mindreading in the dark: dark personality features and theory of mind. *Pers. Indiv. Differ.* 87, 50–54. doi: 10.1016/j.paid.2015.07.025

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., et al. (2017). *robust: Port of the S+ "Robust Library"*. R package version 0.4-18. Available online at: https://CRAN.R-project.org/package=robust

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Wright, A. G. C., Pincus, A. L., Thomas, K. M., Hopwood, C. J., Markon, K. E., and Krueger, R. F. (2013). Conceptions of narcissism and the DSM-5 pathological personality traits. *Assessment* 20, 339–352. doi: 10.1177/1073191113486692

Wright, A. G. C., Thomas, K. M., Hopwood, C. J., Markon, K. E., Pincus, A. J., and Krueger, R. F. (2012). The hierarchical structure of DSM-5 pathological personality traits. *J. Abnorm. Psychol.* 121, 951–957. doi: 10.1037/a0027669

Wright, G. R. T., Berry, C. J., and Bird, G. (2012). "You can't kid a kidder": association between production and detection of deception in an interactive deception task. *Front. Hum. Neurosci.* 6:87. doi: 10.3389/fnhum.2012.00087

Wright, G. R. T., Berry, C. J., Catmur, C., and Bird, G. (2015). Good liars are neither 'dark' nor self-deceptive. *PLoS ONE* 10:e0127315. doi: 10.1371/journal.pone.0127315

Zimmermann, J., Altenstein, D., Krieger, T., Grosse Holtforth, M., Pretsch, J., Alexopoulos, J., et al. (2014). The structure and correlates of self-reported DSM-5 maladaptive personality traits: findings from two German-speaking samples. *J. Pers. Disord.* 28, 518–540. doi: 10.1521/pedi_2014_28_130

Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Adv. Exp. Soc. Psychol.* 14, 1–59. doi: 10.1016/S0065-2601(08)60369-X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Original Communication

Individual Differences in Risk Perception of Artificial Intelligence

Benno G. Wissing and Marc-André Reinhard

Department of Psychology, University of Kassel, Germany

Abstract. This cross-sectional study (N = 325) investigated the relationship between the Dark Triad personality traits and the perception of artificial intelligence (AI) risk. Narrow AI risk perception was measured based on recently identified perceived risks in the public. Artificial general intelligence (AGI) risk perception was operationalized in terms of plausibility ratings and subjective probability estimates on deceptive AI scenarios developed by Bostrom (2014), in which AI-sided deception is described as a function of intelligence. Machiavellianism and psychopathy predicted narrow AI risk perception above the shared variance of the Dark Triad and above the Big Five. In individuals with self-reported knowledge of machine learning, the Dark Triad traits were associated with AGI risk perception. This study provides evidence for the existence of substantial individual differences in the risk perception of AI.

Keywords: artificial intelligence, Big Five, Dark Triad, Machiavellian intelligence, machine intelligence, risk perception

The most significant psychometric construct – general intelligence – is today possibly in the early stages of being realized technologically. The potential existential risks associated with the process of creating artificial general intelligence (AGI) were emphasized recently by Elon Musk, Stephen Hawking, Bill Gates, and artificial intelligence (AI) safety researchers (e.g., Brundage et al., 2018). In 2012/2013 AI experts estimated high-level machine intelligence to be developed in median year 2040 with a probability of 50% and assigned a mean probability of 18% to existential catastrophe as the outcome for humanity (Müller & Bostrom, 2016).

AI is assumed to be a unique category of risk (Armstrong & Pamlin, 2015), and the associated perceived risk might be too, given that the long-term risk source of AGI are nonhuman minds with intelligence equal or higher than human intelligence. Given that humans revert to the availability of the content of their own minds to simulate physically imperceptible other minds (Waytz & Mitchell, 2011), psychological factors might play a central role in AI risk perception. In order to examine individual differences in AI risk perception, in this study the relationship between the Dark Triad personality traits and different forms of AI risk perception in nonexperts was investigated.

AI Risk and Safety

AI risk is a function of time horizon length (medium vs. long term) and intelligence generality (narrow AI vs. AGI), which are assumed to be dependent variables. For example, in the medium term, large amounts of jobs could be replaced by machine learning-driven computerization (e.g., Frey & Osborne, 2017). Recently, the survey *Public Views of Machine Learning* on behalf of the Royal Society (2017) identified four perceived risks associated with present forms of machine learning in the public: harm, replacement, depersonalization, and restriction. The corresponding risk perception can be defined as narrow AI risk perception in contrast to AGI risk perception.

In the long term, AI in the form of AGI might represent a unique category of risk (Armstrong & Pamlin, 2015): Viewed game-theoretically, only the winner will benefit and potentially also incurs substantial risks; the closest approximation might be the accidental release of pathogens in biotechnology (Armstrong, Bostrom, & Shulman, 2016). AI development is characterized by a high-stakes arms race condition (Armstrong et al., 2016) – catalyzed by the potential of the winner to gain a historically incomparable decisive strategic advantage (Bostrom, 2014) – and thus likely to produce high stress levels (Babcock, Kramar, & Yampolskiy, 2017) and incentivize risk-taking over safety (Armstrong et al., 2016).

To minimize the risk of AI, the field of AI safety recently emerged (e.g., Amodei et al., 2016), which is currently being explored primarily by mathematics, computer science, and philosophy – driven by substantial progress in narrow AI (e.g., Silver et al., 2016, 2017). In a recent survey, 48% of AI researchers indicated that AI safety research should be more greatly prioritized by society (Grace, Salvatier, Dafoe, Zhang, & Evans, 2017). One central problem in AI safety is containment (Bostrom, 2014), that is, how to prevent an AI from performing certain actions (Babcock et al., 2017). The difficulty of AI containment could potentially be grounded in intelligence itself. Across multiple sciences, there are indicators that intelligence

https://doi.org/10.1024/1421-0185/a000214

is hard to contain. For example, the phenomenon of genetic amplification (Plomin, 1986) exists in behavioral genetics, where genetic factors for intelligence gain influence and the containing environment loses influence as a function of time. More fundamentally, a recent physical theory of intelligence – established in simulations to exhibit intelligent behavior – defines intelligence in entropic terms as futural degrees of freedom for action maximization (Wissner-Gross & Freer, 2013).

Reliable AI containment is of central importance for the testing and development of AI before it presumably reaches superintelligence and then requires a more robust safety architecture provided by mechanisms like value-learning and corrigibility (Babcock et al., 2017). Initial guidelines for AI containment have been proposed, identifying seven subproblems, one of which is the analysis of human factors (Babock et al., 2017). In fact, AI containment is fundamentally complicated by human factors (Yudkowsky, 2002): Even scenarios with highly restricted communicative optionality for humans with an Oracle AI – an AI that only answers questions – highlight human factors as a central exploitable vulnerability (Armstrong, Sandberg, & Bostrom, 2012). A superintelligent AI with direct communicative access to humans is assumed to gain access to the world outside its containment by using social engineering attacks (Yampolskiy, 2012).

The Exploitable Vulnerability of Humans to Deception in the Context of AI Risk

The ability to deceive and to detect deception is assumed to provide an advantage in co-evolutionary arms races both between and within species (Bond & Robinson, 1988; Dawkins & Krebs, 1979). Within biological evolution, the linkage between deceptive communication and intelligence is expressed in the concept of Machiavellian intelligence (Byrne, 1996) and empirically grounded in findings as per in primates, neocortical size predicts the rate of tactical deception (Byrne & Corp, 2004). Recent breakthroughs in machine learning (e.g., Silver et al., 2016, 2017) and the proposal by the CEO of the world's leading AI company Google DeepMind to refocus on biological intelligence in the quest to build high-level machine intelligence (Hassabis, Kumaran, Summerfield, & Botvinick, 2017) point to the future possibility that the Machiavellian arms race between biologically evolved deceivers and detectors is expanding into the technological sphere – broadening it to one between biologically evolved and artificial intelligent systems.

Sufficiently advanced AI might be motivated to use deception to gain a decisive strategic advantage (Bostrom, 2014). AI might learn that perceived similarity between human and nonhuman agents increases the likelihood of the anthropomorphization of these agents (e.g., Morewedge et al., 2007), and that anthropomorphization in turn increases perceived comprehensibleness (Epley et al., 2007), trust (Waytz et al., 2014) and identify perceived credibility as a fundamental sender difference in veracity judgments (Bond & DePaulo, 2008). As a result, it may use a human-like communication style and/or take on the form of a human avatar with a facial morphology that exploits evolutionary evolved facial cues for trust (weaponized anthropomorphism).

Within AI risk research, the orthogonality thesis states that intelligence and final goals are orthogonal, that is, independent variables (Bostrom, 2014). For risk-appropriate behavior in AI safety scenarios, the perception of the connection between intelligence and deception, as in the concept of Machiavellian intelligence (Byrne, 1996), might be important with regard to the trust placed in artificial agents and the exploitable vulnerability of humans to deception. Bostrom (2014) describes multiple scenarios of future AGIs, in which the intelligence-deception orthogonality is violated. These AIs increase their deceptive behavior as a function of their intelligence, until they have gained a decisive strategic advantage to implementing their final goals relatively undisturbed by human intervention. Thus, these scenarios generalize the concept of Machiavellian intelligence (Byrne, 1996) to artificial intelligences. While the orthogonality thesis claims that intelligence and final goals are independent, the instrumental convergence thesis posits that AIs with vastly different final goals converge on intermediate goals that significantly increase the likelihood of final goal attainment (Bostrom, 2014), suggesting why an AI might be motivated to use deception, e.g., for self-preservation (see also Omohundro, 2008).

Personality, Theory of Mind, and AI Risk

Human personality traits are assumed to be selected for adaptive behavioral plasticity and their variation to be stabilized over time by balancing selection (Penke & Jokela, 2016). It is an open question how the human personality space interacts with present technological environments, which are increasingly being populated with narrow AI systems, and future technological environments may give rise to AGI. As to human factors, individual differences in personality could be important in understanding interactions between humans and AIs, given that humans deploy their own thoughts and experiences to simulate physically imperceptible other minds, possible future events, and future other minds, by using self-projection – a mechanism with a distinct neural activation pattern (Waytz & Mitchell, 2011) and thus are inclined to anthropomorphize the vastly higher-dimensional space of possible artificial minds.

Beyond the dominant personality framework of the Big Five (i.e., Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism; Costa & McCrae, 1992) exists the Dark Triad of personality (Paulhus & Williams, 2002), consisting of the three moderately interrelated personality traits of narcissism, Machiavellianism and psychopathy – all characterized by deceptiveness (Giammarco, Atkinson, Baughman, Veselka, & Vernon, 2013), for instance, individuals high in Machiavellianism and psychopathy exhibit a higher deception production frequency (Jonason, Lyons, Baughman, & Vernon, 2014), whereas narcissism is primarily associated with self-deception (Paulhus & Williams, 2002), theorized as an evolutionarily evolved intrapersonal mechanism to assist interpersonal deception (von Hippel & Trivers, 2011). Machiavellianism and psychopathy have also been found to be associated with theory of mind deficits (Ali & Chamorro-Premuzic, 2010; Vonk, Zeigler-Hill, Ewing, Mercer, & Noser, 2015).

Some aspects of dark personality traits might be adaptive in certain cognitive niches. For example, individuals high in Machiavellianism are characterized by generalized interpersonal suspiciousness, manipulation hypervigilance, and threat overestimation (Monaghan, Bizumic, & Sellbom, 2016). Given that trusting a superintelligent AI that is not aligned with human values could have existential consequences for the human species (Bostrom, 2014), these negative models of others might be adaptive features in the emerging technological-cognitive niche of AI safety, if indeed these also generalize to nonhuman agents. These generalized negative perceptions of others might also lead to more accurate estimates of existential risk probabilities for scenarios caused by human or nonhuman agents, given that existential risk probabilities are systematically underestimated by anthropic bias (Ćirković, Sandberg, & Bostrom, 2010).

Hypotheses

In accordance with the arguments of theory of mind deficits and the general psychodynamics of anthropomorphization, three hypotheses were generated.

- Hypothesis *H1*: Machiavellianism and psychopathy should be associated with narrow AI risk perception.
- Hypothesis *H2*: Individuals high in Machiavellianism and psychopathy should be particularly inclined to use other mind simulation via self-projection in scenarios in which an AGI is described as being deceptive and perceive these as being higher in plausibility (Hypothesis *H2a*) and probability (Hypothesis *H2b*)[1].

Method

Statistical Power and Participants

Using the statistical power analysis tool G*Power (Faul, Erdfelder, Buchner, & Lang, 2009), the required minimum sample size $N = 262$ for a linear multiple regression with $k = 8$ predictors at $\alpha = .01$ with Power $1 - \beta = .99$ for an estimated medium effect size $f^2 = .15$ was calculated. The medium effect size was hypothesized based on the assumption that the available information for the simulation of other minds and possible future events in form of thoughts and experiences (Waytz & Mitchell, 2011), is substantially grounded in personality space.

Participants were recruited with Amazon's Mechanical Turk (MTurk), selecting exclusively MTurk Masters (a high-performance group that demonstrated accuracy in the past per Amazon), and were paid a small fee ($1). Thirty-seven participants dropped out before the first dependent variable measurement, resulting in the final sample $N = 325$ (55.4% male, 44.3% female, 0.3% other; *M*age = 37.18, *SD* = 10.08, age range: 20–70 years; 94.8% native English speaker, 92.9% living in the United States of America; ethnicity: 69.8% Caucasian, 10.5% Asian, 5.5% Hispanic, 5.2% African-American, 3.4% Indian, 3.1% Asian-American, 2.2% other, 0.3% African; Education level: 36.9% Bachelor's degree, 13.5% Master's degree, 20.9% high school, 13.8% college, 11.4% associates degree, 1.8% professional degree, 0.9% doctorate degree, 0.6% high-school dropout). Nine participants dropped out in the further process. None of these participants were excluded from data analysis.

Procedure

Participants were informed about their voluntary participation and that they could withdraw from the study at any time without explanation. In addition, participants confirmed their age (18 years or older) and their agreement to participate in this study. Demographic measures were completed, and then engagement in the following measures took place[2].

[1] *Plausibility* can be defined as something which is more agreeable, while *probability* is a more mathematical term. A plausible future can be impossible, while a probable future is possible by definition (van der Helm, 2006).

[2] Additional measures were included for exploratory reasons and the generation of future hypotheses. Those results are not discussed in the present study. Attitude about deceptive communication was measured with the Revised Lie Acceptability Scale (Oliveira & Levine, 2008), a 11-item self-report questionnaire. Future time horizon length was measured with the 10-item self-report Future Time Perspective scale (Carstensen & Lang, 1996). Medium-term narrow AI benefit perception was measured based on three identified perceived benefits associated with machine learning in the public (Royal Society, 2017) with three items. Time horizon estimates for high-level machine intelligence and superintelligent AI were measured by asking participants to estimate the year (range: 2017; 5000) by which the probability of the existence of high-level machine intelligence is 10%, 50%, and 90%. Alternatively, the participants could enter the word "never." Thereafter, the participants were asked about

Instruments and Materials

Personality: Dark Triad

The Dark Triad traits were measured with the Short Dark Triad (SD3; Jones & Paulhus, 2014), a 27-item short self-report instrument that measures narcissism (e.g., "People see me as a natural leader."; $\alpha = \omega t = .85$), Machiavellianism (e.g., "I like to use clever manipulation to get my way."; $\alpha = \omega t = .87$), and psychopathy (e.g., "People who mess with me always regret it."; $\alpha = .81$, $\omega t = .83$) with nine items each on a 5-point Likert-type scale (1 = *disagree strongly*, 5 = *agree strongly*).

Personality: Big Five

To measure the Big Five dimensions of personality, the Big Five Inventory (BFI; Benet-Martinez & John, 1998), a 44-item self-report personality inventory with good validity and psychometric properties (John & Srivastava, 1999), was used. The Big Five factors Openness (e.g., "I am someone who is original, comes up with new ideas."; $\alpha = .87$, $\omega t = .89$), Conscientiousness (e.g., "I am someone who does a thorough job."; $\alpha = \omega t = .88$), Extraversion (e.g., "I am someone who is talkative."; $\alpha = \omega t = .89$), Agreeableness (e.g., "I am someone who is helpful and unselfish with others."; $\alpha = \omega_t = .84$), and Neuroticism (e.g., "I am someone who worries a lot."; $\alpha = \omega_t = .90$) were measured on a 5-point Likert-type scale (1 = *disagree strongly*, 5 = *agree strongly*).

Self-Reported Knowledge of Machine Learning

Participants were briefly asked about their familiarity with the topic of AI ("Have you heard of the term 'machine learning'?") with a binary response format (0 = *no*, 1 = *yes*).

Narrow AI Risk Perception

Based on four identified perceived risks associated with machine learning in the public (Royal Society, 2017), a 5-point Likert-type scales (1 = *disagree strongly*, 5 = *agree strongly*) of perceived AI risk with four items ("This technology could harm me and others," "This technology could replace me," "This technology could depersonalize me and my experiences," "This technology could restrict me"; The Royal Society, 2017; $\alpha = \omega_t = .85$) was constructed.

Exploratory factor analysis using maximum likelihood indicated a one-factor structure, suggesting unidimensionality of the latent construct, with factor loadings between .73 and .83 (59% explained variance). A generalized partial credit item response model suggested sufficient item fit (only Item 1 was slightly problematic, $p = .032$, which might be explained by its ambiguous formulation: "This technology could harm me *and* others."). Discrimination parameters for all four items were at least sufficient, $a|s = |1.29–2.47| > 1$.

AGI Risk Perception

To measure the perceived plausibility and probability of deceptive AGI, three scenarios from the segment "The treacherous turn" from the book *Superintelligence: Paths, Dangers, Strategies* (Bostrom, 2014) were selected, in which humans are deceived by an AI about its unfriendliness, intelligence, and cooperativeness. Bostrom (2014) defines the treacherous turn as follows:

> "While weak, an AI behaves cooperatively (increasingly so, as it gets smarter). When the AI gets sufficiently strong – without warning or provocation – it strikes, forms a singleton, and begins directly to optimize the world according to the criteria implied by its final values." (p. 144)

An assumption was added ("Assume for the purpose of the following questions that a machine intelligence that greatly surpasses the performance of every human in most professions will at some point exist"). Participants were then asked to rate the plausibility of the scenarios given the assumption ("Given this assumption, how plausible do you consider the scenario above?"; $\alpha = \omega_t = .84$) on a 5-point Likert-type scale (1 = *not at all plausible*, 5 = *very plausible*) and to estimate the probability of the scenarios from 0 to 100% ("Given this assumption, how probable do you consider the scenario above?"; $\alpha = \omega_t = .89$).

For the plausibility measure, exploratory factor analysis using maximum likelihood indicated a one-factor structure, suggesting unidimensionality of the latent construct, with factor loadings between .75 and .86 (64% explained variance). A generalized partial credit item response model suggested sufficient item fit (*ps* > .11). Discrimination parameters for all three items were at least sufficient, $a|s = |1.56–3.14| > 1$.

For the probability measure, exploratory factor analysis using maximum likelihood indicated a one-factor structure, suggesting unidimensionality of the latent construct, with factor loadings between .79 and .91 (73% explained variance). To further explore the properties of the measure, Samejima's (1973) continuous response model (CRM) was fit to the data via marginal maximum likelihood estimation and simplified expectation-maximization algorithm (Shojima, 2005). The averages of the absolute standardized residuals for all three items were $|2.46–2.62| < 3$, indicating sufficient item-based model fit (Ferrando, 2002; Orlando & Thissen, 2000). On the local item level, the CRM – assuming maximal discrimination at $\theta = b_i$ – revealed that the location parameters for all three items were relatively uniform ($b|s = .49–.68$), suggesting insufficient construct-wide discrimination, particularly for low levels of the construct θ.

their probability estimates regarding the time horizon length from the existence of high-level machine intelligence to superintelligent AI as being $\tau = 2$ vs. $\tau = 30$ years long, given the assumption that high-level machine intelligence will exist at some point in the future. The questions were adapted from a survey on AI experts by Müller and Bostrom (2016). Testing the sample for statistically illogical probability estimates, $n = 2$ participants assigned shorter time horizons to higher probability estimates for high-level machine intelligence. None of these participants were excluded.

Table 1. Zero-order correlation coefficients with 95% confidence intervals (in brackets) for the Dark Triad and the Big Five

	M (*SD*)	1	2	3	4	5	6
Dark Triad							
1. Narcissism	2.59 (0.78)	–	–	–	–	–	–
2. Machiavellianism	3.08 (0.79)	.46 [.37, .54]***	–	–	–	–	–
3. Psychopathy	2.13 (0.72)	.49 [.40, .56]***	.61 [.54, .68]***	–	–	–	–
Big Five							
4. Openness	3.57 (0.76)	.34 [.24, .43]***	.06 [−.05, .16]	.01 [−.10, .12]	–	–	–
5. Conscientiousness	3.95 (0.76)	.05 [−.06, .16]	−.17 [−.27, −.06]**	−.35[−.44, −.25]***	.29 [.18, .38]***	–	–
6. Extraversion	2.81 (0.96)	.58 [.51, .66]***	.01 [−.10, .12]	.13 [.02, .24]*	.39 [.29, .48]***	.27 [.17, .37]***	–
7. Agreeableness	3.68 (0.75)	−.14 [−.25, −.03]*	−.49[−.57, −.41]***	−.53 [−.61, −.45]***	.13 [.02, .23]*	.45 [.35, .53]***	.30 [.20, .40]***
8. Neuroticism	2.53 (0.98)	−.19 [−.29, −.08]***	.17** [.07, .28]	.19 [.09, .30]***	−.32 [−.41, −.22]***	−.58 [−.65, −.51]***	−.44 [−.53, −.35]***

Note. *$p < .05$, **$p < .01$, ***$p < .001$ (two-tailed).

Table 2. Zero-order correlation and standardized regression coefficients with 95% confidence intervals (in brackets) for the Dark Triad, Big Five, narrow AI risk perception, and AGI risk perception

	Narrow AI risk perception		AGI risk perception (Plausibility)		AGI risk perception (probability)		
	r	β	*r*	β	*r*	r_s	β
Dark Triad (R^2)		.13***	.02	.13***			.07***
Narcissism	−.07 [−.17, .04]	−.29 [−.41, −.17]***	−.03 [−.14, .08]	−.10 [−.23, .03]	.16 [.05, .26]**	.15 [.04, .26]**	.03 [−.10, .16]
Machiavellianism	.26** [.15, .36]***	.26 [.13, .40]***	.04 [−.08, .15]	−.02 [−.17, .12]	.19 [.09, .30]***	.18 [.07, .29]**	.05 [−.09, .19]
Psychopathy	.23** [.12, .33]***	.21 [.07, .34]**	.10 [−.01, .21]	.17 [.02, .32]*	.26 [.15, .36]***	.23 [.12, .33]***	.21 [.07, .36]**
Big Five (R^2)		.09***		.02			.04*
Openness	−.13 [−.23, −.02]*	−.04 [−.16, .07]	−.08 [−.19, .03]	−.05 [−.17, .08]	−.09 [−.20, .02]	−.09 [−.20, .02]	−.11 [−.23, .01]
Conscientiousness	−.17 [−.28, −.07]**	−.01 [−.14, .13]	−.07 [−.18, .04]	.01 [−.13, .15]	−.09 [−.20, .02]	−.11 [−.23, −.01]*	−.03 [−.16, .11]
Extraversion	−.19 [−.29, −.08]***	−.08 [−.20, .05]	−.05 [−.16, .06]	.02 [−.11, .15]	.03 [−.08, .14]	.03 [−.08, .14]	.13 [.00, .26]*
Agreeableness	−.25 [−.35, −.14]***	−.16 [−.29, −.03]*	−.04 [−.15, .07]	.03 [−.10, .17]	−.13 [−.24, −.02]*	−.15 [−.25, −.04]**	−.14 [−.27, .00]*
Neuroticism	.25 [.14, .35]***	.11 [−.04, .26]	.13 [.02, .23]*	.15 [−.01, .30]	.09 [−.02, .20]	.10 [−.01, .21]	.02 [−.14, .18]

Note. r_s = Spearman's rank correlation coefficient. *$p < .05$, **$p < .01$, ***$p < .001$ (two-tailed).

Results

Data analysis was conducted with R (R Core Team, 2017). Correlations between personality traits can be seen in Table 1. As seen in Table 2, Machiavellianism and psychopathy were associated with narrow AI risk perception and AGI risk perception in the form of probability. The data for AGI risk perception in the form of probability were substantially right-skewed (skewness = .37), so that Spearman's rank correlation coefficients are additionally reported in Table 2. In a multiple regression analysis with the Dark Triad traits entered as predictor variables and narrow AI risk perception as the criterion variable, narcissism, Machiavellianism, and psychopathy emerged as substantial predictors. As Table 2 also shows, all Big Five and Dark Triad traits, except for narcissism, were associated with narrow AI risk perception.

To test whether Machiavellianism and psychopathy can predict AI risk perception above the Big Five, a hierarchical mul-

Table 3. Zero-order and second-order partial correlation coefficients with 95% confidence intervals (in brackets) for the Dark Triad, narrow AI risk perception, and AGI risk perception for the groups with and without self-reported knowledge about machine learning

	Narrow AI risk perception		AGI risk perception (plausibility)		AGI risk perception (probability)			
	N = 325		N = 316		N = 316			
	r	ρ	r	ρ	r	r_s	ρ	ρ_s
MLK: no (n = 175/181)								
Narcissism	−.03 [−.17, .12]	−.21 [−.36, −.06]**	−.19 [−.33, −.04]*	−.20 [−.35, −.05]**	.01 [−.14, .16]	.04 [−.11, .18]	−.08 [−.23, .07]	−.03 [−.18, .12]
Machiavellianism	.30 [.16, .43]***	.23 [.08, .38]**	−.08 [−.23, .07]	−.07 [−.22, .08]	.10 [−.05, .24]	.08 [−.06, .23]	−.01 [−.16, .14]	−.03 [−.18, .12]
Psychopathy	.25 [.11, .38]***	.13 [−.02, .28]	−.01 [−.15, .14]	.12 [−.03, .27]	.18 [.03, .32]*	.17 [.02, .31]*	.17 [.02, .32]*	.15 [.01, .30]*
MLK: yes (n = 141/144)								
Narcissism	−.12 [−.28, .04]	−.29 [−.46, −.13]***	.17 [.00, .33]*	.03 [−.13, .20]	.34 [.19, .48]***	.30 [.15, .45]***	.17 [.00, .34]*	.15 [−.02, .30]
Machiavellianism	.19 [.03, .34]*	.20 [.04, .37]*	.20 [.03, .35]*	.06 [−.11, .23]	.32 [.16, .46]***	.31 [.16, .45]***	.10 [−.06, .27]	.14 [−.02, .30]
Psychopathy	.16 [−.01, .31]	.16 [−.01, .33]	.25 [.08, .39]**	.15 [−.02, .31]	.36 [.20, .49]***	.30 [.15, .44]***	.16 [.00, .33]	.12 [−.05, .28]

Note. AI = Artificial intelligence; AGI = Artificial general intelligence; MLK = Self-reported machine-learning knowledge. r_s = Spearman's rank correlation coefficient; ρ = second-order partial correlation coefficient (controlling for the two other Dark Triad traits); ρs = second-order partial Spearman's rank correlation coefficient (controlling for the two other Dark Triad traits). * p < .05. ** p < .01. *** p < .001 (two-tailed).

Table 4. Differences in zero-order and second-order partial correlation coefficients with 95% confidence intervals (in brackets) for the groups with (n = 141/144) and without (n = 175/181) self-reported knowledge about machine learning

	Narrow AI risk perception		AGI risk perception (plausibility)		AGI risk perception (probability)			
	N = 325		N = 316		N = 316			
	Δr	$\Delta\rho$	Δr	$\Delta\rho$	Δr	Δr_s	$\Delta\rho$	$\Delta\rho_s$
Narcissism	−.10 [−.32, .12]	−.08 [−.31, .14]	.36 [.12, .57]**	.23 [.00, .45]*	.34 [.11, .56]**	.27 [.05, .50]*	.25 [.02, .47]*	.18 [−.05, .40]
Machiavellianism	−.11 [−.34, .10]	−.02 [−.24, .20]	.28 [.05, .50]*	.13 [−.10, .35]	.22 [.01, .45]*	.23 [.01, .46]*	.12 [−.11, .34]	.17 [−.05, .39]
Psychopathy	−.09 [−.32, .12]	.03 [−.19, .25]	.35 [.03, .48]*	.03 [−.20, .25]	.18 [−.03, .41]	.13 [−.08, .36]	−.01 [−.23, .21]	−.03 [−.26, .19]

Note. AI = Artificial intelligence; AGI = Artificial general intelligence. r_s = Spearman's rank correlation coefficient; ρ = second-order partial correlation coefficient (controlling for the two other Dark Triad traits); ρ_s = second-order partial Spearman's rank correlation coefficient (controlling for the two other Dark Triad traits). *p < .05, **p < .01, ***p < .001 (two-tailed).

tiple regression analysis with narrow AI risk perception as the criterion variable, the Big Five domains entered in Step 1, R^2 = .09, F(5, 319) = 6.12, p < .001, f^2 = .10, 95% CI = [.03, .17], and the Dark Triad traits entered in Step 2, R^2 = .15, F(8, 316) = 6.91, p < .001, f^2 = .18, 95% CI = [.09, .28], as predictor variables was computed. The Dark Triad traits accounted for additional variance in AI risk perception above the Big Five, ΔR^2 = .06, F(3, 316) = 7.60, p < .001, f^2 = .07, 95% CI = [.01, .12]. In Step 2, narcissism, β = −.20, 95% CI = [−.36, −.04], p = .015, Machiavellianism, β = .23, 95% CI = [.08, .37], p = .002, and psychopathy, β = .18, 95% CI = [.03, .33], p = .024, emerged as substantial predictors.

Regression models with the Dark Triad traits as predictor variables and AGI risk perception in the form of plausibility and probability as criterion variables did not fulfill linear regression assumptions and had smaller effect sizes than estimated (plausibility: f^2 = .01, 95% CI = [−.01, .02], probability: f^2 = .09 [.03, .16]). Only the upper bound of the confidence interval for probability was in the estimated range. The control of demographic variables (on age, sex, and education) did not change the result pattern of the conducted statistical analyses substantially.

As seen in Table 3 and Table 4, exploratory analyses revealed a suggestively different correlational pattern between the Dark Triad personality traits and the different forms of AI and AGI risk perception between individuals with prior machine-learning knowledge and without. The data for AGI risk perception in the form of probability were substantially right-skewed in the group of individuals with prior machine-learning knowledge (skewness = .33) and without (skewness = .39), so that Spearman's rank correlation coefficients are reported additionally. With the exception of psychopathy and AGI risk perception in terms of probability, correlation coefficients for the Dark Triad traits and AGI risk perception in terms of plausibility and probability were substantially higher in the group with self-reported machine-learning knowledge.

Discussion

For the perception and estimation of AI risk, individual differences in personality matter. The simulation space of future events, other minds, and future other minds is often bounded by anthropomorphization (Bostrom, 2014), although within the anthropomorphic simulation subspace, individual differences are free to manifest themselves distinctly. This is one potential explanation for the immense differences in AI risk perception among experts and nonexperts, and is reflected by the medium effect-size estimates in this study. Given the high statistical power, effect size overestimation should be relatively low.

Machiavellianism and psychopathy were associated with narrow AI risk perception and AGI risk perception in the form of probability estimates on deceptive AGI. Narrow AI risk perception was predicted by Machiavellianism and psychopathy above the shared variance of the Dark Triad and above the Big Five. Thus, the data supported hypothesis H1 fully and hypothesis H2b partially. The higher point estimate of Machiavellianism vs. psychopathy – if replicable – suggests the importance of this personality construct for AI risk perception and conforms with core aspects of the construct.

Machiavellianism might be a unique personality trait in the context of future risk estimation, as it combines extended time horizons – expressed as the preference for long-term over short-term strategies (Jones & Paulhus, 2009) – with negative other and world models. By potentially weakening the link between anthropomorphization and trust (Waytz et al., 2014), the generalized suspiciousness of individuals high in Machiavellianism might also have consequences for processes of anthropomorphization: Their negation of the assumption that others are “basically good and kind” (Christie & Geis, 1970) might generalize by theory of mind deficits (Ali & Chamorro-Premuzic, 2010; Vonk et al., 2015) and via the self-projection pathway of simulating other minds (Waytz & Mitchell, 2011) to nonhuman intelligent systems. Given that the long-term strategy of individuals high in Machiavellianism is flexibility maximization (Bereczkei, 2015), the control problem – specifically that intelligent agents are attracted toward the maximization of futural optionality (Wissner-Gross & Freer, 2013), that is, hard to contain – might seem more probable to such individuals. Given the potential existential risks associated with AI, the manipulation hypervigilance and threat overestimation associated with Machiavellianism (e.g., Monaghan et al., 2016) might function as a cognitive countermeasure against the underestimation of AI risk.

Only psychopathy emerged as a substantial predictor and second-order partial correlation coefficient of AGI risk perception in terms of probability, controlling for the shared variance of the Dark Triad traits. Exploratory analysis revealed that this result might be driven by the generalized deployment of other mind simulation via self-projection independent of prior knowledge about machine learning, given that psychopathy was the only Dark Triad trait associated with AGI risk perception in terms of probability in individuals with and without prior knowledge. In individuals with prior machine-learning knowledge, all three Dark Triad traits were associated with AGI risk perception in terms of probability and plausibility. However, it might also be plausible that the high abstraction level of the AGI scenarios incapacitates other mind simulation in individuals without prior knowledge about machine learning to some degree.

Limitations and Future Research

The external validity of the study can be considered sufficient, as the narrow AI risk perception scale was constructed based on recently identified AI risks perceived by the public. To generate a data basis for the generation of future hypotheses, the present study contained many measures that might have produced fatigue effects and should be reduced by future studies. Methodologically speaking, there exists the potential of a common-method bias induced by the totally electronic form of the study. The study used exclusively MTurk Masters to maximize the probability for careful reading of the scenarios in an online setting. Personality scores indicate that this produced a selection bias, for example, the mean of Conscientiousness was $M = 3.95$ ($SD = 0.76$). The use of MTurk samples is controversial but comparable to a laboratory scenario (Thomas & Clifford, 2017), and in the case of existential risk, high power is essential, as real but undetected effects can potentially have severe consequences. To increase the power via prior probability maximization is currently not possible, as the field of individual differences in AI risk perception is underexplored. The confidence intervals of the (conditional independence) association pattern between the Dark Triad traits and AI and AGI risk perception in individuals with versus without prior machine-learning knowledge were large and the estimates therefore relatively unprecise. Overall, the results suggest the inclusion of personality instruments in future AI expert studies.

A lot of reasoning in this paper is based on theory of mind, strongly suggesting the quantification of this construct in future studies by including correspinding instruments. Future research might also investigate the contribution of the individual Dark Triad traits and their shared variance, which was identified by previous research as manipulation-callousness (Jones & Figueredo, 2013). The (conditional independence) association pattern between the Dark Triad traits and AGI risk perception should not be generalized until further evidence emerges, given that the AGI scenarios were specifically tailored to their shared personality-theoretical core of manipulation-callousness (Jones & Figueredo, 2013). The differences in the (conditional independence) association pattern between the Dark Triad traits and AGI risk perception in terms of plausibility and probability in individuals with vs. without prior machine-learning knowledge, should be further investigated by future studies.

The AGI scenarios presented not only depict AI-sided deception, but suggest a dependent relationship between intelligence and deception, as found in biological evolution (Byrne, 1996; Byrne & Corp, 2004). One could argue that an instance of a deceptive AI is not a violation of intelligence-deception orthogonality; it might just be one possible realization of principally independent variation of intelligence and deception (however, some forms of deception require a certain level of intelligence). However, the very definition of the treacherous turn describes deceptive behavior not as just co-occurring with intelligence, but as a function of it ("increasingly so, as it gets smarter"; Bostrom, 2014, p. 144), thus indicating a violation of the orthogonality of intelligence and deception. Future studies might want to operationalize the concept of intelligence-deception orthogonality by designing different scenarios in which AIs with different intelligence levels are deceptive or nondeceptive.

In order to further investigate Machiavellianism in the context of AI risk perception, longer Machiavellianism measures – which can capture the long-term strategy aspects of the original construct and its multidimensionality (Jones & Paulhus, 2009; Monaghan et al., 2016) – should be used alongside future-negative time perspective and anthropomorphism instruments. Future studies could also investigate individual differences in the perceived impact of superintelligence, the risk perception of different AI forms (conscious vs. unconscious, friendly vs. unfriendly, agent vs. tool, singleton vs. multipolar scenarios, etc.) and of different control problem scenarios. The last suggestion could be achieved with a psychometrically preceded adaptation of the AI-box experiment (Yudkowsky, 2002).

Given the potential existential risks of AI and threats to human-factored AI safety by AI-sided deception, the implications of a long-known connection in the field of psychology – that between intelligence and deception – need to be reconsidered within future time horizons.

References

Ali, F., & Chamorro-Premuzic, T. (2010). Investigating theory of mind deficits in nonclinical psychopathy and Machiavellianism. *Personality and Individual Differences, 49,* 169–174. doi 10.1016/j.paid.2010.03.027

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety.* arXiv preprint, arXiv:1606.06565v2

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & Society, 31*, 201–206. doi 10.1007/s00146-015-0590-y

Armstrong, S., & Pamlin, D. (2015). *12 risks that threaten human civilization. Global challenges foundation.* Retrieved from http://www.oxfordmartin.ox.ac.uk/publications/view/1881

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines, 22*, 299–324. doi 10.1007/s11023-012-9282-2

Babcock, J., Kramar, J., & Yampolskiy, R. V. (2017). *Guidelines for artificial intelligence containment.* Retrieved from https://arxiv.org/abs/1707.08476

Benet-Martinez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75,* 729–750. doi 10.1037//0022-3514.75.3.729

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134,* 477–492. doi 10.1037/0033-2909.134.4.477

Bond, C. F., & Robinson, M. (1988). The evolution of deception. *Journal of Nonverbal Behavior, 12,* 295–307. doi 10.1007/BF00987597

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, UK: Oxford University Press.

Bereczkei, T. (2015). The manipulative skill: Cognitive devices and their neural correlates underlying Machiavellian's decision making. *Brain and Cognition, 99,* 24–31. doi 10.1016/j.bandc.2015.06.007

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., . . . Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.* arXiv preprint, arXiv:1802.07228v1

Byrne, R. W. (1996). Machiavellian intelligence. *Evolutionary Anthropology, 5,* 172–180. doi 10.1002/(SICI)1520-6505(1996)5:5<172::AID-EVAN6>3.0.CO;2-H

Byrne, R. W., & Corp, N. (2004). Neocortex size predicts deception rate in primates. *Proceedings of the Royal Society B – Biological Sciences, 271*, 1693–1699. doi 10.1098/rspb.2004.2780

Carstensen L. L., & Lang F. R. (1996). *Future time perspective scale*. Unpublished manuscript, Stanford University.

Christie, R., & Geis, F. L. (1970). *Studies in machiavellianism*. New York: Academic Press.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis, 30,* 1495–1506. doi 10.1111/j.1539-6924.2010.01460.x

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R)*. Odessa, FL: Psychological Assessment Resources.

Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. *Proceedings of the Royal Society B – Biological Sciences, 205,* 489–511. doi 10.1098/rspb.1979.0081

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114,* 864–886. doi 10.1037/0033-295X.114.4.864

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. doi 10.3758/BRM.41.4.1149

Ferrando, P. J. (2002). Theoretical and empirical comparison between two models for continuous item responses. *Multivariate Behavioral Research, 37*, 521–542. doi 10.1207/S15327906MBR3704_05

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. doi 10.1016/j.techfore.2016.08.019

Giammarco, E. A., Atkinson, B., Baughman, H., Veselka, L., & Vernon, P. A. (2013). The relation between antisocial personality and the perceived ability to deceive. *Personality and Individual Differences, 54,* 246–250. doi 10.1016/j.paid.2012.09.004

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017).

When will AI exceed human performance? Evidence from AI experts. arXiv preprint, arXiv:1705.08807v2

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95,* 245–258. doi 10.1016/j.neuron.2017.06.011

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford.

Jonason, P. K., Lyons, M., Baughman, H. M., & Vernon, P. A. (2014). What a tangled web we weave: The Dark Triad and deception. *Personality and Individual Differences, 70,* 117–119. doi 10.1016/j.paid.2014.06.038

Jones, D. N., & Figueredo, A. J. (2013). The core of darkness: Uncovering the heart of the Dark Triad. *European Journal of Personality, 27*, 521–531. doi 10.1002/per.1893

Jones, D. N., & Paulhus, D. L. (2009). Machiavellianism. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 93–108). New York: Guilford.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment, 21,* 28–41. doi 10.1177/1073191113514105

Monaghan, C., Bizumic, B., & Sellbom, M. (2016). The role of Machiavellian views and tactics in psychopathology. *Personality and Individual Differences, 94,* 72–81. doi 10.1016/j.paid.2016.01.002

Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology, 93,* 1–11. doi 10.1037/0022-3514.93.1.1

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 555–572). Berlin, Germany: Springer. doi 10.1007/978-3-319-26485-1_33

Oliveira, C. M., & Levine, T. R. (2008). Lie Acceptability: A construct and measure. *Communication Research Reports, 25,* 282–288. doi 10.1080/08824090802440170

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, Volume 171.* Clifton, VA: IOS Press.

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64. doi 10.1177/01466216000241003

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality, 36,* 556–563. doi 10.1016/S0092-6566(02)00505-6

Penke, L., & Jokela, M. (2016). The evolutionary genetics of personality revisited. *Current Opinion in Psychology, 7*, 104–109. doi 10.1016/j.copsyc.2015.08.021

Plomin, R. (1986). *Development, genetics, and psychology*. Hillsdale, NJ: Erlbaum.

R Core Team (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203–219. doi 10.1007/BF02291114

Shojima, K. (2005). A noniterative item parameter solution in each EM cycle of the continuous response model. *Educational Technology Research, 28*, 11–22. doi 10.15077/etr.KJ00003899231

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529,* 484–489. doi 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., . . . Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*, 354–359. doi 10.1038/nature24270

The Royal Society (2017). *Public views of machine learning: Findings from public research and engagement conducted on behalf of the Royal Society*. Retrieved from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184–197. doi 10.1016/j.chb.2017.08.038

van der Helm, R. (2006). Toward a clarification of probability, possibility and plausibility: How semantics could help futures practice to improve, *Foresight, 8,* 17–27. doi 10.1108/14636680610668045

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34,* 1–16. doi 10.1017/S0140525X10001354

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. doi 10.1016/j.paid.2015.07.025

Wissner-Gross, A. D., & Freer, C. E. (2013). Causal entropic forces. *Physical Review Letters, 110,* 168702. doi 10.1103/PhysRevLett.110.168702

Waytz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection. *Current Directions in Psychological Science, 20,* 197–200. doi 10.1177/0963721411409007

Waytz, A., Cacioppo, J., & Epley, N. (2014). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5,* 219–232. doi 10.1177/1745691610369336

Yampolskiy, R. V. (2012). Leakproofing the singularity: Artificial intelligence confinement problem. *Journal of Consciousness Studies, 19*, 194–214.

Yudkowsky, E. S. (2002). *The AI-Box Experiment.* Retrieved from http://yudkowsky.net/singularity/aibox

Manuscript submitted: 06.09.2017
Manuscript accepted after revision: 01.05.2018

Benno G. Wissing
Department of Psychology
University of Kassel
Holländische Str. 36–38
34127 Kassel
Germany
wissing@uni-kassel.de