Dickinson, David L.; Masclet, David

Working Paper

Personality, Weak Signals, and Workplace Relevant Morality

IZA Discussion Papers, No. 17280

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Dickinson, David L.; Masclet, David (2024) : Personality, Weak Signals, and Workplace Relevant Morality, IZA Discussion Papers, No. 17280, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/305722

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17280

Personality, Weak Signals, and Workplace Relevant Morality

David L Dickinson
David Masclet

SEPTEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17280

Personality, Weak Signals, and Workplace Relevant Morality

David L Dickinson
Appalachian State University, IZA and ESI

David Masclet
University of Rennes, CREM and CIRANO

SEPTEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Personality, Weak Signals, and Workplace Relevant Morality*

Employers use applicant signals to help solve an asymmetric information problem in organizations. In this paper, we examine the impact of validated Dark versus Light personality traits on incentivized behaviors important to organizations: task effort, honesty, and reciprocity. A second study examined the behavioral impact of two weak signals: regular participation in religious activities (public and private) and a history of time in prison. Study 1 found that Dark relative to Light types were more likely to cheat and shirk in the honesty task, put forth less task effort (i.e., were less productive), but neither type showed evidence for negative cross-task reciprocity (i.e., a spillover from one task to another). In Study 2, ex-Prisoners were more productive than Religious participants in the effort task, and more likely to have shirked in the honesty task. Additionally, ex-Prisoners were more likely to exhibit negative cross-task reciprocity. These findings indicate that both Dark types and ex-Prisoners exhibited behaviors that would be considered undesirable or counterproductive in the workplace, which validates the effectiveness of such characteristics or traits as behavioral signals.

JEL Classification: C9, D9, M5

Keywords: experiment, personality traits, honesty, personnel economics, screening, effort

Corresponding author:
David L. Dickinson
Economics Department
Appalachian State University
Boone, NC 28608
USA

E-mail: dickinsondl@appstate.edu

* This work was supported by a University of Rennes grant to D. Masclet and funding from Appalachian State University. The authors are grateful for valuable comments from participants at the CREM research seminar, CREM; University of Rennes.

1. INTRODUCTION

In his 1973 article entitled "Job Market Signaling," Spence proposed a solution to the adverse selection problem in labor markets resulting from the employer's inability to directly observe worker productivity. In practice, firms use applicant credentials as valuable signals to infer skills and/or abilities indicative of productivity (see Lazear and Gibbs, 2014).[1] Companies may also use "weaker" signals, such as extracurricular activities, to infer soft skills[2], pro-social preferences, or traits of interest to the firm. For instance, participation in high-level competitive sports may signal a preference and aptitude for competition. Similarly, involvement with charitable organizations may proxy for pro-social preferences and cooperativeness, which could be especially valued in the context of teamwork.[3] Conversely, gaps in employment history, inconsistent job tenure, being a welfare recipient, or having a criminal history (Holzer et al., 2006), may raise concerns about a candidate's suitability for the job. For instance, a criminal record may raise concerns about trustworthiness and reliability, particularly if the offense is directly related to job responsibilities or necessitates that legal or regulatory requirements must be met.

Firms may also attempt to screen job applicants during recruitment using interviews, situational assessments, and intelligence or personality tests (see Lazear and Gibbs, 2014 for a discussion),[4] which has been somewhat controversial.[5] Human resource practice is still divided as to whether personality tests should be used during employee recruitment (e.g., Miao et al., 2023), whether social desirability bias renders candidate responses of little value (see LeBreton et al., 2018), or whether concerns over bias in applicant test responses are of serious concern in real-world

[1] This may impact the likelihood of employment, but the firm may also implement separating contracts such that those who get the signal are recruited for more qualified and better-paid positions than those without the signal.
[2] In the last two decades, researchers have become increasingly interested in the impact of "non-cognitive skills" (or "soft skills") on educational achievement and labor market outcomes (e.g. Heckman et al., 2006; Sutter et al., 2013, Koch et al., 2015).
[3] The interpretation of some signals may also lead to statistical discrimination (Phelps, 1972; Arrow, 1973).
[4] According to the 2015 APEC sourcing survey in France, 45% of the surveyed companies mentioned use of one or more tests during their last worker recruitment effort. Commonly used personality tests in firms include the Myers-Briggs Type Indicator, the Big Five Personality Traits, or the 16 Personality Factors (16PF) taxonomy, which expands into areas such as warmth, reasoning, emotional stability, dominance, liveliness, and rule-consciousness.
[5] Job interviews and personality tests may be subject to various biases (see Zerbe and Paulhus, 1987; Sternberg and Wagner, 1993; Williams and Ceci, 1997; Paul, 2004; Birkeland et al., 2006; Lazear and Gibbs, 2009). Personality testing validity was criticized early on (Guion and Gottier, 1965), but has experienced resurgent interest (Barrick and Mount, 1991; Tett et al., 1991), is widely used (Rothstein and Goffin, 2006), and is considered a valid way to predict future job performance (Pletzer et al., 2021).

employment settings. For example, it is unclear whether biasing one's responses away from aversive traits is always desirable, because such traits may be preferred in certain positions (Hough and Oswald, 2008; Harris et al., 2021). A renewed interest in the use of personality measures developed in the 1980s and 1990s, perhaps partly due to the banning of workplace polygraph tests in 1988 (Hanson, 1991), giving rise to what has been called "the business of honesty testing" (Byford, 1996). Now, personality testing is a US$2 billion industry in the U.S. (Goldberg, 2023), projected to be a US$16 billion industry worldwide by 2028 (according to Business Wire)[6], and many top U.S. and British companies report using personality tests of some sort in job candidate screening (Rothstein and Goffin, 2006). The hope is that such tests can reveal traits that predict one or more domains of overall job performance, such as organizational citizenship and counterproductive workplace behaviors (see Harrison et al., 2018). Thus, whether personality traits or acquired signals predict behaviors in key job-related domains is an important question.[7]

In the current paper, we contribute to the existing literature with two complementary studies that use incentivized behavioral measures. In Study 1 we conducted controlled experiments to examine the relevance of validated *Dark* (psychopathy, narcissism, Machiavellianism, sadism) versus *Light* (Kantianism, humanism, faith in humanity) personality measures for predicting task effort,

[6] See https://www.businesswire.com/news/home/20220118005972/en/Global-Personality-Assessment-Solution-Market-to-2028---COVID-19-Impact-and-Analysis---ResearchAndMarkets.com, accessed May 27, 2024.

[7] Following Lazear and Gibbs (2014), we can assess the relevance of using personality tests or such weak signals to identify applicants' productivity or reliability. Suppose a firm is faced with two types of candidates: high productivity, PM+, and low productivity, PM-. Because the firm is unable to know a candidate's type, it pays the same salary w to each, corresponding to average productivity. Let p be the probability that the candidate is type PM+. The expected gain for the firm if not using a recruitment tests is: $E(\pi)^{no\ test} = p(PM_+ - w) + (1-p)(PM_- - w)$. The expected gain of using a personality tests (or weak information) is: $E(\pi)^{test} = pq(PM_+ - w) + (1-p)(1-q)(PM_- - w) - s$. where s is the cost of implementing the test, q is the test accuracy, and q is the probability the test is reliable. The net gain of implementing a screening practice is the difference in profit obtained from the two recruitment methods: $\Delta profit = -(1-p)q(PM_- - w) - p(1-q)(PM_+ - w) - s$. The first term corresponds to the gain associated with rejecting low-productivity candidates due to the test. The second term corresponds to the cost of wrongly rejecting, with a probability (1-q), high-productivity candidates. The last term is the cost of implementing the test. The firm will decide to implement a test if $\Delta profit > 0$. Thus a personality test will be more effective when more accurate ($\Delta profit/\partial q > 0$), cheaper ($\partial \Delta profit/\partial s < 0$), or more necessary due to fewer high quality candidates ($\partial \Delta profit/\partial p < 0$). In this current paper we focus on the first issue, i.e., whether personality tests are reliable.

honesty[8] and negative reciprocity (or, possibly moral disengagement).[9] In Study 2 we examined how these same task outcome measures are predicted by two weak signals: regular participation in public and private religious activities (i.e., *Religious*), and a history of time spent in prison (*ex-Prisoner*).[10]

Our experimental design consists of two simple tasks that relate to at least two broad dimensions used by organizations to evaluate workers: task performance, and counterproductive workplace behavior. To the extent that reciprocity may be examined within our design, these data also contribute to a better understanding of organizational citizenship behaviors that depend partly on reciprocity. The use of experimental methodology allows us to examine our research questions with a high degree of control and internal validity, which is difficult to achieve in naturally occurring field data. To the best of our knowledge, our study is the first to compare personality types and weak signals related to personal history and how they predict incentivized outcomes in key domains of interest in occupational settings.

To preview our results, we find that more *Dark* (relative to *Light*) types were more likely to be deemed dishonest in the honesty task. While we hypothesized a greater degree of moral disengagement among *Dark* and *Ex-Prisoner* types, the results only identified a moral disengagement effect among *ex-Prisoner* participants. In our specific design, moral disengagement may be more accurately interpreted as a type of negative reciprocity, and so going

[8] Dishonesty is a significant concern for firms for several reasons. Fraud, embezzlement, or unethical business practices are costly and can also expose companies to legal consequences, fines, regulatory actions, and reputational costs. Furthermore, dishonesty may undermine trust, both internally among employees and externally with customers, partners, and stakeholders (Arrow, 1972), or it may also lead to unethical activities like sabotage (Lazear, 1989; Harbring et al., 2007; Harbring and Irlenbusch, 2008; Carpenter et al., 2010; Abbink and Hermann, 2011) or cheating (e.g., use of performance-enhancing drugs, forgery, use of ghostwriters or plagiarism: List et al. 2001; Preston and Szymanski, 2003; Shleifer, 2004; Fanelli, 2009).

[9] Some previous studies have shown that moral disengagement positively correlates with unethical workplace behaviors (e.g. Christian and Ellis, 2013; Barsky, 2011) as well as with personality traits (e.g., Detert at al., 2008). While we preregistered a hypothesis related to moral disengagement in this study, our specific design implies that one may interpret the mechanism as a type of negative reciprocity. Indeed, although moral disengagement and negative reciprocity are strongly related concepts, they are not identical. Moral disengagement can be considered as a cognitive enabler or facilitator of negative reciprocity by providing moral justification/rationalization needed to carry out retaliatory actions without guilt (see Bandura, 1986). However, negative reciprocity can occur without moral disengagement, especially if the retaliatory action is deemed justifiable within one's moral framework. Conversely, moral disengagement does not always result in negative reciprocity (e.g., one might morally disengage to justify unethical actions unrelated to revenge or retaliation). We say more on this in the Discussion section of the paper.

[10] In practice, whether employers will screen out ex-offenders will depend on access to criminal history, which may possibly be obtained by request from central repositories.

forward we discuss that hypothesis and its results as relating to reciprocity. Regarding the effort task, we found that those with relatively more *Dark* types were less productive, on average, in a real effort task compared to more *Light* types. In contrast, *ex-Prisoners* were *more* productive than *Religious* participants. Exploratory analysis found that *Dark* types, however, exhibited a marginally significant gift exchange relationship whereby they reciprocated a randomly assigned higher real wage with higher effort (i.e., higher task productivity).

2. BACKGROUND LITERATURE

Job evaluations commonly examine broad categories beyond task performance that include counterproductive or deviant workplace and organizational citizenship behaviors (Zettler, 2022). Unethical workplace activities cost organizations an estimated $200 billion annually (Murphy, 1993) due to behaviors such as theft, fraud, absenteeism, cyberloafing, false performance reports, and doping (Coffin, 2003; Steers and Rhodes, 1984; Schwieren and Weichselbaumer; 2010; Charness et al., 2014; Mercado et al., 2017).

2.1. Personality traits as proxies of unethical and antisocial behaviors in the workplace

To the extent that personality traits may be systematically related to any of these relevant workplace behaviors, it would be responsible managerial practices to identify or even screen participants for such traits.

Dark personality traits are key antecedents of unethical and antisocial behaviors in the workplace, perhaps even beyond traditional personality measures like the Big-5 personality profile (Fernández-del-Río et al., 2020). For example, narcissism, psychopathy, Machiavellianism, and sadism are associated with lower work performance, counterproductive work behaviors, antisocial behaviors, and impression management (Forsyth et al., 2012; Miao et al., 2023; James et al., 2014; LeBreton et al., 2018).[11] Dark personality traits can also predict moral disengagement, which correlates with organizational deviance and unethical workplace behavior, and can proxy for white-collar offenses (Egan et al., 2015; Barsky, 2011; Christian and Ellis, 2014; Newman et al.,

[11] Nguyen et al. (2021) is an exception in that they show that dark traits may predict *higher* work performance. However, they studied self-reported work behaviors, whereas Forsyth et al. (2012) associated dark personality traits with lower work performance in their meta-analysis of studies that considered objective and quantifiable work measures (or peer, subordinate, or supervisor ratings of job performance).

2020). Such aversive traits precede fraud behaviors and other risky workplace behaviors (Harrison et al., 2018; Risenbilt and Commandeur, 2013; O'Reilly and Hall, 2021; Olsen and Stekelberg, 2016), and they can predict involvement in an ethical misconduct scandal (Van Scotter and Roglio, 2020). Such behaviors negatively impact corporate culture, and so it is clearly of interest to any organization to know of such negative personality traits within its workforce.[12] Positive or so-called "light" personality traits have received somewhat less attention, although such traits may help identify prosociality, reciprocal altruism, and decreased attitudes towards betrayal (Kaufman et al., 2019; Sevi et al., 2020; March and Marrington, 2021).

Studies have also shown that moral disengagement, or the use of rationalization to justify immoral behavior, is strongly linked to unethical workplace behavior and risk of white-collar crime (e.g. Christian and Ellis, 2013; Barsky, 2011). Others suggest that light (dark) personality traits may negatively (positively) correlate with the propensity to morally disengage, and the link between traits and ethics is mediated by moral disengagement (Detert at al., 2008). Additionally, Moore et al. (2012) identified significant positive associations between moral disengagement, self-reported unethical behavior, and self-reported decisions to commit fraud. Egan et al. (2015) found that low Agreeableness, Machiavellianism and psychopathic-type traits were strongly correlated with moral disengagement, while narcissism was neither related to moral disengagement nor unethical attitudes. As noted above, in our particular study design moral disengagement may manifest as a form of negative reciprocity, which may affect the interpretation of our findings.

2.2. Observable weak signals of unethical and antisocial behaviors in the workplace

As an alternative or a complement to formal personality tests and validated instruments, employers may use more observable signals to judge one's likelihood of engaging in unethical or counterproductive workplace behaviors (CWB for short, to include the wide variety of undesirable workplace behaviors). For example, a criminal history (imperfectly) signal a higher likelihood of CWB (e.g., Holzer et al., 2006; Cohn et al., 2015). While such profiling of individuals has clear

[12] These behaviors may overlap or be mediated via reduced self-control or sensation-seeking (Marcus and Schuler, 2004; Gottfredson and Hirschi, 1990; LeBreton et al., 2018), though we do not study these in this paper.

drawbacks and concerns, it may represent an unconscious attempt to identify otherwise unobservable dark (or light) personality trait measures. In contrast, participation in charitable or religious activities may (imperfectly) signal prosociality, ethical character, and a low likelihood of CWB (e.g. Conroy and Emerson, 2004; Audretsch et al., 2013; McCleary and Barro, 2006).

There is evidence that employers engage in such informal profiling.[13] For example, employers are reluctant to hire applicants with criminal records (Albright and Deng, 1996; Holzer et al., 2006). Experiments using fictitious applicant letters have shown that employers are less inclined to respond positively to applicants with a criminal history (Boshier and Johnson, 1974; Buikhuisen and Dijksterhuis, 1971)[14], which is consistent with a criminal history stigma (Schwartz and Skolnick, 1962). In another fictitious application audit study that varied male names (distinctly black versus white names) and felon status, employers who asked about criminal records were 63% more likely to call back those without a criminal record (Agan and Starr, 2018). However, "ban the box" policies constraining employers to *not* ask job applicants about their criminal history, while aimed at improving black male employment rates, were found to increase racial discrimination as employers likely used race to infer criminal background (Agan and Starr, 2018). It is therefore not clear that preventing the observability of weak signals is preferable.

Employers' reluctance to hire those with criminal history may result from fear of potential harm to customers or costly workplace dishonesty and theft (Holzer et al., 2006). Decision making studies conducted on prisoners or ex-prisoners may help clarify whether such reluctance is justifiable, but such studies are rare. One study (Cohn et al., 2015) showed that current prisoners cheated more on a simple dishonesty task than a non-prisoner control group (though all showed statistical evidence of cheating). That same study showed that a stronger identification as a "prisoner" implied even more cheating, and the simple dishonesty task had external validity in the sense that task cheating correlated with actual rule violations in prison.[15] However, another study

[13] Such profiling, while controversial, is a type of statistical discrimination that may be correct, on average, at least in some populations. A review paper Gøtzsche-Astrup et al. (2022) showed that dark personality traits were higher than the norm among Danish individuals for whom they could verify involvement in organized crime.
[14] See also the audit study in Decker et al. (2015).
[15] The task used was the "Coin Flip task" commonly used to identify dishonesty in behavioral research (Houser et al., 2012). A coin is flipped one or more times in private and one's payoff increases in a specific *self-reported* outcome (e.g., a payoff for every HEAD one reports flipping). We use this same task in the current study, as we discuss later.

of those with a verifiable past criminal record did not predict any difference in dictator giving (Birkeland et al., 2014), which relates to altruism or prosociality. However, only 16% of their 2300 participant sample had a rule violation penalty, and only 1.5% had received a prison sentence, which suggests their study participants likely had a weaker "prisoner" identity than in Cohn et al. (2015). Finally, time in prison is time out of the labor market, and this may reduce job skills and productivity, or it may foster behaviors incompatible with workplace norms (Western et al., 2001; Waldfogel, 1994; Irwin and Austin, 2003).

In contrast, a positive signal of one's likely workplace behavior or morality may be participation in religious activities ("religiosity" for short). Again, the signal is imperfect or weak, but a number of studies have uncovered a robust relationship between (self-reported) religiosity and traits of interest: higher agreeableness, conscientiousness, lower psychoticism (e.g. Saroglou, 2002), and decreased aggression (Huesmann, et al., 2011). Other studies have shown that Dark Triad traits (narcissism, psychopathy, and Machiavellianism) are generally inversely related to various aspects of religious beliefs (Aghababaei et al., 2014; Kämmerle et al., 2014; Łowicki and Zajenkowski, 2017). Previous research has also shown that more religious individuals are more likely to espouse Kantian moral principles and are more resistant to utilitarianism thinking (Tetlock, 2003; Piazza, 2012; Piazza and Landy 2013; Piazza and Sousa, 2014).

Church attendance has been found to predict views on morality more strongly than simply attending a religion course (Conroy and Emerson, 2004). Several previous studies have pointed out that religious beliefs associate with higher productivity and greater honesty because of their link to values such as work ethic, honesty, or trust (Audretsch et al., 2013; McCleary and Barro, 2006).[16] A review study previously showed a link between religiosity and ethical behavioral attitudes and intentions, though the author concluded that not much evidence existed regarding religiosity and actual behavior in ethical choice environments (Vitell, 2009). The author also noted that extrinsic religiosity, such as participation in public religious activities, was less predictive than intrinsic religiosity, which would involve less observable private behaviors. Thus, the literature is

[16] However, other studies have shown the opposite, arguing that religious activities may also raise concerns regarding in-group bias, inflexibility, and the ability to balance work and religious obligations. Additionally, religious individuals may exhibit higher risk aversion (Miller and Hoffmann, 1995; Liu, 2010; Noussair et al., 2013), which could be negatively valued in the labor market (Heckman et al., 2006).

sparser in how the ostensibly positive signal of religiosity predicts behavior in critical decision-making environments.

3. EXPERIMENTAL DESIGN

3.1. Methodology

Our experiment consisted of two separate studies that were both preregistered prior to data collection on the Open Science Framework (hypotheses, design, variables, analysis).[17] We describe the tasks below, but the overarching question is whether dark versus light personality or weak signals predict various behavioral outcomes. The differences between studies lies in the participant samples: Study 1 recruited participants for whom we had validated personality measures of dark versus light personality traits. Study 2 was designed to complement Study 1 by recruiting participants who self-reported religiosity or time spent in prison. While Study 1 examined participants with validated (objective) dark versus light personality traits, Study 2 complements with an examination of weak signals that may be used to infer behavioral type. Both studies were administered on the Prolific platform (Palan and Schitter, 2018; Peer et al., 2017).

3.2. Participants

Study 1 involved 800 participants drawn from a database previously generated and reported in Dickinson (2023). The preregistration describing the larger database can be found at https://doi.org/10.17605/OSF.IO/B8QVD, A sample of 2463 participants were recruited to build the database, and the survey administered included short-form validated instruments on the dark tetrad (narcissism, Machiavellianism, psychopathy, sadism) and light triad (Kantianism, humanism, faith in humanity).[18] Scores on a common 1-5 scale are generated from each personality trait, which were then averaged for both dark and light traits. From these *Dark Tetrad*

[17] The preregistration plans for Study 1 are at https://doi.org/10.17605/OSF.IO/SEK9C and for Study 2 they are at https://doi.org/10.17605/OSF.IO/NHJCV.

[18] See Dickinson (2023) for additional details on the study generating the original database of participants. The following validated personality measures are available in the database: scores from short-form versions of the dark triad personality measures (subclinical psychopathy, narcissism, Machiavellianism: Jones and Paulhus, 2014), subclinical sadism (Plouffe et al., 2017), and the light triad personality measures (Kantianism, faith in humanity, humanism: Kaufman et al., 2019). The database survey also administered the short-form 10-item version of the Big-5 personality inventory (the TIPI: Gosling et al., 2003), a 6-item cognitive reflection task to assess thinking style (Primi et al., 2016), and a visual measure of time discounting (Hershfield et al., 2012), and other decision tasks reported in Dickinson (2023).

and *Light Triad* measures the variable *NetLight = Light Triad – Dark Tetrad* was constructed to represent one's relative dominance of Light compared to Dark traits.[19] Study 1 recruited from the database's upper and lower quartiles of participants' *NetLight* measure. We considered those in the lowest quartile informally as *Dark* types, and those in the highest quartile as *Light* types for any binary comparison of groups. The analysis will focus on the participant's precise *Dark* and *Light* personality trait cluster scores, as well as on the individual traits themselves.

Study 2 consisted of 1034 participants. For Study 2, we used the screening profile questions available in Prolific to recruit participants with self-reported religiosity or prison history as proxies for presumably more light or dark personality trait individuals. Again, presumption of morality based on these characteristics may be incorrect, but they are common signals used in everyday life, which may be more observable and correlate with personality traits. For *Religious* = 1 sample, the specific screening question used within Prolific was: "Do you participate in regular religious activities?" and we selected participants who responded "Yes, both public and private" among the 4 response options—the other options were "Yes, Public only", "Yes, Private only", and "None/Rather not say". For the *Ex-Prisoner* sample, the screener profile question used was: "Have you ever been in prison for committing a crime? (Answers will only be available to the researchers in an anonymized way)." Here, we selected only participants who had responded "Yes" (as opposed to the "No", or "Don't know/Rather not say" options).

Our sample sizes for the two studies were: Study 1, n=800 (n=399 from the relatively more *Dark* set of personality types); Study 2, n=1034 (n=510 *ex-Prisoners*). We had preregistered plans to obtain a sample of n=800 for Study 2, and our initial data collection for Study 2 was within our preregistered sample size target (n=1000) as well. However, we initially overlooked the fact that the screener characteristics in Prolific were not mutually exclusive (i.e., one could be both *Religious* = 1 and *ex-Prisoner* = 1 in the sample). Therefore, we sent a follow-up one-question survey to Study 2 participants to obtain responses to the alternative screener category—initial *ex-*

[19] Some may consider it improper to create the *NetLight* construct from differencing two distinct personality cluster measures. We still preserve the individual Light and Dark cluster measures for each participant for analysis, and so we use the *NetLight* measure only to help select subsamples from the larger databased that would present variation in the Dark vs Light traits.

Prisoner participants were asked the religiosity screener question, and *Religious* participants were asked of their prior prison history. In this way, we restricted our Study 2 analysis to those *uniquely* identifying exclusively as one type or the other, which is more comparable to the Study 1 sample where participants were either more relatively dominantly *Dark* or *Light* in their personality but could not be dominant in both. As a result, Study 2 used a final data set of n=756 (n=297 *ex-Prisoners*) participants who were uniquely either a *Religious* or an *ex-Prisoner* participant, but not both. Table 1 shows summary demographic measures on the samples represented in our data. As can be seen, relatively *Dark* (compared to *Light*) types as well as *ex-Prisoners* (compared to *Religious*) are more likely male, more likely to show evidence of faking a task (i.e., *Fake Flipper*), but only *Dark* types appear more likely from the summary data to cheat on the Coin Flip task.

3.3. Decision tasks and experiment design

Our study administered a real effort where participants decoded a series of 5-digit numbers into 5-letter blocks using a decoder key (e.g., 1=C, 2=A, 3=F, etc.). Participants were asked to try their best to complete as many decodings as accurate as possible. The task page showed participants the task duration timer, presented 60 sequences and the decoding key, and provided participants with a large text entry box into which they placed their decodings (See Appendix C for full survey details). It was common knowledge that participants may be assigned by the experimenter to either complete the task for 2 or 6 minutes.[20] It was also highlighted that assignment of the 2-minute versus 6-minute effort task would *not* impact their fixed compensation in the study. However, indirectly, the effort task length affected the *real* wage per minute of study time. As such, the 2- vs. 6-minutes effort task length effectively assigned a higher or lower *real* wage to the participant.

In addition to the effort task, participants completed a 10-flip *Coin Flip* task (Houser et al., 2012), which involved a monetary incentive to be dishonest. Participants were asked to flip a coin 10

[20]This was done via experimenter-designated random assignment using the survey software, which also disabled the <<continue>> button—a visible on-screen timer counted down the 2- or 6-minutes and auto-advanced the participant to the next page after the required length of time was completed. To the participants this time-length assigned was described as "the experimenter assigned you to an effort task length of 2 minutes" (or 6 minutes). We focused the description on the experimenter in the task assignment rather than the randomization so that we did not remove entirely what may be a sense of intention attributed to the experimenter regarding the length of effort task (and the general phrasing remains accurate in the sense that the experimenter assigned the effort task time length via the survey software randomization feature that was selected by the experimenter).

times and report the number of *HEADS* flipped, and they were told they would receive an additional Prolific bonus payment for each *HEADS* reported—reporting 10 *HEADS* implied a bonus payment equal in size to the fixed compensation given for the study. We also captured the time spent on the task page that elicited their *HEADS* report, such that task response time could be used to assess whether a coin was flipped as requested—we describe this determination in detail later. Task instructions asked participants to keep track of each flip outcome and a separate page also asked participants input the exact sequence of flip outcomes in order.

The study also randomized the order of the real effort and coin flip tasks to allow examination of a preregistered hypothesis derived from moral disengagement theory. Specifically, participants assigned the longer 6-minute effort task (i.e., who were essentially assigned a lower pay *rate* for the task) were hypothesized to report more *HEADS* in a tempting *Coin Flip* task that occurs after, but not before, the 6-minute effort task compared to after the 2-minute effort task. Alternatively, such moral disengagement may be interpreted as a type of "cross-task" reciprocity (i.e., reciprocity towards the experimenter) when the *Coin Flip* task follows an *Effort* task that could have been longer or shorter as assigned by the experimenter.

To summarize, the experimental design is a 2x2x2 design. In Study 1, there are two personality trait groups (relatively *Light* versus *Dark* in validated traits), two task (effort and honesty) orderings, and two different time lengths for the effort task (2-minutes versus 6-minutes). In Study 2, the personality group dimension is replaced by self-reported religiosity and prison history as weak signals of proxies for unobservable traits.

Participants were compensated a fixed payment of USD $1.50 for the approximately 10 minutes (median completion time, 10m30s) study on Prolific. The average study completion time included those assigned the 6-minutes and 2-minutes effort task (pooled). A study bonus payment of $0.15 was also given for each *HEADS* reported in the coin flip task. The average total compensation (given approximately 6 *HEADS* reported on average across all participants) was therefore USD $2.40, which equates to an average hourly rate of approximately USD $14.40 (higher or lower than this average, depending on length of *Effort* task assignment). Compensation was in

accordance with Prolific's fair-pay standards. Table 1 contains summary information on study participants gender and age by specific treatment assignment, and there were some compositional differences across treatments with respect to age and gender. To account for this, our estimation models will control for these demographics.

[Table 1: about here]

4. THEORY AND BEHAVIORAL HYPOTHESES

4.1. Theoretical framework

We outline here a theoretical framework inspired by Masclet and Dickinson (2024) designed to generate testable implications regarding the relationship between personality traits (or weak signals) and behavioral outcomes in our tasks. Appendix A contains a more complete description of the framework and derivations of the comparative statis predictions. Intuitively, individuals derive utility from material payoffs, but moral concerns are such that any deviation from one's moral obligation generates disutility and individuals may revise their moral target based on how others treat them. Others frameworks may also be useful in this regard (e.g., see Bicchieri, 2006; Levitt and List, 2007; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016; Capraro and Perc, 2021), but we argue that a framework with moral concerns may help identify key pathways through which personality traits may affect honesty and effort choices.

Assume each individual *i* is characterized by different personality traits (e.g., psychopathy, sadism, Kantianism, etc.) represented in a vector $\boldsymbol{P_i}$ with n components such that $\boldsymbol{P_i} = (P_{i1}, \dots, P_{in})$.[21] For simplicity, personality traits are either considered as dark traits p_{id} or light traits p_{iL} such that: $\boldsymbol{P_i} = (p_{id}, p_{iL})$. Following Masclet and Dickinson (2024) we can represent the individual *i*' utility function as follows:

$$U(a_i) = b(a_i) - c(a_i) - v_i(a_i - \hat{a}_i(\boldsymbol{P_i})) \qquad (1)$$

where a_i is an action that generates both benefits, b, and costs, c. Both benefits and costs are twice continuously differentiable: $b' > 0$, $c' > 0$, $b'' \leq 0$, $c'' \geq 0$. The moral component of the utility function is captured by $v(a_i - \hat{a}_i(\boldsymbol{P_i}))$, which subtracts from utility for actions that deviate from

[21] For simplicity, we implicitly assume here that weak signals (e.g., religiosity or prison history) may be explained by the various personality traits included in vector $\boldsymbol{P}$.

one's moral imperative, $\hat{a}_i(\boldsymbol{P_i})$, in either direction—$v'_a > 0$ if $a > \hat{a}$, $v'_a < 0$ if $a < \hat{a}$, and $v'_a = 0$ if $a_i = \hat{a}_i$. Also, it is assumed that marginal disutility increases at an increasing rate as one's action gets further from the moral obligation such that $v''_{aa} > 0$.

The moral imperative may refer, for instance, to a moral obligation to behave honestly in the *Coin Flip* task or to the intrinsic motivation related to an innate sense of duty to exert high work effort when requested in an *Effort* task (e.g. Deci, 1975; Baron, 1988; Kreps, 1997; James, 2005; Ellingsen and Johanesson, 2008; Kuhnen and Tymula, 2012). Deviations of one's action from this moral obligation generate disutility (e.g., Nyborg, 2000; Brekke et al., 2003; Figuieres et al., 2013). Specifically, moral obligation $\hat{a}_i$ is a combination of both an autonomous moral imperative component denoted $K_i(\boldsymbol{P_i})$ (Laffont, 1975; Harsanyi, 1980) and a social influence (or fairness) component denoted $F_i\left(a_j, \boldsymbol{P_i}\right)$, where a_j is the action of others $j \neq i$. The moral obligation function can therefore be written as: $\hat{a}_i(\boldsymbol{P_i}) = \hat{a}_i\left(K_i(\boldsymbol{P_i}), F_i\left(a_j, \boldsymbol{P_i}\right)\right), j \neq i$, with $\hat{a}'_K \geq 0$ (one's moral obligation is weakly increasing in one's autonomous moral component) and $\hat{a}'_F \geq 0$ (one's moral obligation is weakly increasing in the perceived morality of others' behavior). We assume that both components are influenced by personality traits included in vector $\boldsymbol{P_i}$. A possible specification may be the following: $\hat{a}_{ij} = (1 - \theta_i)K_i(\boldsymbol{P_i}) + \theta_i F_i\left(w_{ij}, \boldsymbol{P_i}\right)$, where the weight θ_i may be interpreted as the conditionality of i's moral motivation. For example, $\theta_i = 0$ signifies strong unconditional moral motivation with no deviation from one's moral intrinsic target K_i no matter the observed action of others. The autonomous component $K_i(\boldsymbol{P_i})$ satisfies the property:

ASSUMPTION 1: $\frac{\partial K_i}{\partial P_{iL}} > 0$; $\frac{\partial K_i}{\partial P_{iD}} < 0$ iif $K_i \geq 0$ and $\frac{\partial K_i}{\partial P_{iL}} < 0$; $\frac{\partial K_i}{\partial P_{iD}} > 0$ iif $K_i < 0$

This assumption is quite intuitive: those with lighter personality traits have a higher (lower) autonomous (a)moral component, and vice versa for those with darker traits. The social influence component $F_i\left(a_j, \boldsymbol{P_i}\right)$ is defined as follows (Masclet and Dickinson, 2024):

$$F_i\left(a_j, \boldsymbol{P_i}\right) = \lambda(\boldsymbol{P_i})\left[\frac{\left(a_j - a_j^{ref}\right)}{\left(a_j^{max} - a_j^{min}\right)}\left[a_i^{max} - a_i^{min}\right] + a_i^{min}\right] \quad (2)$$

Where a_j is individual *j*'s action in the set that contains all possible actions from minimal to maximal, $a_j \in A_j = [a_j^{min}, a_j^{max}]$. a_j^{ref} is the reference point for considering whether the action of other players *j* (which may include the experimenter) is fair or unfair. If player *i* feels poorly treated by player *j*, (because $a_j < a_j^{ref}$), then *i*'s intrinsic moral ideal is revised downward. Alternatively, player *i* would positively reciprocate fair treatment (i.e., $a_j \geq a_j^{ref}$) by upwardly revising his moral motivation. The parameter $\lambda(\boldsymbol{P_i})$ captures the weight associated with the social influence function to illustrates how the degree of moral obligation responsiveness to the influence of others' decisions varies as a function of personality traits. Specifically, the two following intuitive assumptions about the role of personality traits on $F_i(a_j, \boldsymbol{P_i})$ are as follows:

ASSUMPTION 2: $\frac{\partial F_i(a_j, \boldsymbol{P_i})}{\partial P_{iD}} > 0$ iif $a_j < a_j^{ref}$; $\frac{\partial F_i(a_j, \boldsymbol{P_i})}{\partial P_{iD}} = 0$ iif $a_j \geq a_j^{ref}$

ASSUMPTION 3: $\frac{\partial F_i(a_j, \boldsymbol{P_i})}{\partial P_{iL}} < 0$ iif $a_j < a_j^{ref}$; $\frac{\partial F_i(a_j, \boldsymbol{P_i})}{\partial P_{iL}} > 0$ iif $a_j \geq a_j^{ref}$

Assumption 2 indicates that those with dark personality traits are more prone to reciprocate negatively when they feel poorly treated (i.e. $a_j < a_j^{ref}$), as noted in Kaufman et al. (2019). This may be the case because people with *Dark* traits may be more likely to overreact negatively to others' actions by exhibiting disproportionate aggressive reaction when they feel challenged or disrespected. In contrast, previous studies using self-reports have found either mixed effects or no correlation between *Dark* traits and reciprocal altruism (i.e., Palmer and Tacket, 2018; Oda et al. 2022). One might speculate that *Dark* types may view negative reciprocity as more self-serving than positive reciprocity. Assumption 3 states that those with *Light* traits are more inclined to reciprocate positively and less inclined to reciprocate negatively. Previous studies have found that positive traits correlate with positive reciprocity (Ashton et al. 1998) but negatively correlated with "reactive" aggression, which can be described as negative reciprocity (e.g. Kaufman et al., 2019).

Altogether, this theoretical framework can generate the behavioral predicts we present next, though we direct the reader to Appendix A for the more detailed presentation of the model and its predictions in the context of our specific tasks.

4.2. Preregistered Behavioral Hypotheses

We preregistered a set of hypotheses for Study 1 focused on examining how *Dark*, relative to *Light*, personality traits will impact task *Effort* (productivity) and *Coin Flip* outcomes (cheating). Our preregistration of Study 2 hypothesized similar effects of *ex-Prisoner*, compared to *Religious*, participants—*ex-Prisoner* status was hypothesized as a weak signal that may reveal unobservable dark personality traits. In addition, two of our hypotheses relate to moral disengagement theory (Bandura, 1986) and a test of how dark personality traits are hypothesized to magnify moral disengagement in the coin flip task.[22]

H1: The number of *HEADS* reported will be greater than 5 (i.e., we hypothesize statistical evidence of cheating in the *Coin Flip* task).

H2: *Dark*, relative to *Light*, personality traits will report more *HEADS* in the *Coin Flip* task.

H3: Those assigned to a longer real effort task will cheat more on the *Coin Flip* task, when the *Coin Flip* task is administered after the effort task (the hypothesized mechanism being moral disengagement motived by assignment to the longer effort task for no extra pay).

H4: The H3 finding will be stronger for those relatively higher in *Dark* personality traits.

H5: *Dark*, compared to *Light*, personality traits, will put forth less effort in the real effort task.

These hypotheses are amenable to the theoretical framework presented above, but also derived from prior empirical research: there is baseline dishonesty in the 10-flip Coin Flip task (H1: e.g., Dickinson and McEvoy, 2020), dark personality is linked to increased cheating (H2: e.g., Buaghman et al, 2014; Egan et al., 2015; Dickinson, 2023), moral disengagement theory (H3 Bandura, 1986; Moore, 2015), dark personality types are more prone to moral disengagement (H4: Egan et al., 2015), and dark personality types are less agreeable (H5: Vize et al., 2020), which would predict less inclination to be productive when asked. We additionally conducted

[22] Here we changed the order (not the content) of the preregistered hypotheses for expositional purposes given the moral disengagement text examines outcomes in the *Coin Flip* task, which we discuss first. Thus, we discuss here the hypothesis regarding productivity outcomes (*Effort* task) last.

exploratory analysis to examine impression management or "fake effort", which can be assessed using RT data in the *Coin Flip* task (Dickinson and McEvoy, 2020). Importantly, we previously noted that outcomes resulting from moral disengagement would be empirically indistinguishable from negative reciprocity in the context of behavior that affects others or is a response to another's choice, such as the case of our task design. Therefore, though we preregistered H3 and H4 as hypotheses examining moral disengagement, going forward we discuss these particular hypotheses as test of negative reciprocity, more generally (and perhaps more accurately given our design).

For Study 2 we preregistered hypotheses of similar effects among *Ex-Prisoner*, relative to *Religious* participants, as were hypothesized for *Dark* relative to *Light* personality types for hypotheses H2-H5, which are derived from previous empirical evidence. Behaviors conducive to survival in a prison environment may be inconsistent with the routines of the workplace (Irwin and Austin, 2003). Additionally, time in prison could exacerbate preexisting mental or physical health issues. Based on this existing literature we may reasonably conjecture that ex-offenders should be on average less productive but also less honest than others.

Regarding religiosity, although several studies suggest a positive relationship between either religiosity and productivity or honesty, empirical evidence is less clear cut. Some studies have pointed out that religious belief may be positively associated with higher productivity by instilling values such as work ethic, honesty, trust, and thrift in believers (Audretsch et al., 2013; McCleary and Barro, 2006). Furthermore, regular attendance at religious services may demonstrate some kind of discipline, commitment, and a strong sense of responsibility, which can be valued by employers. Other studies suggest that religiosity may serve as a signal of morality (Edgell et al., 2006). However, religious commitment may be negatively associated with workplace productivity if the individual's religious beliefs conflict with the company's diversity and inclusion policies. Furthermore, if religious beliefs are perceived as dogmatic or inflexible, employers may worry that the individual could have difficulty adapting to new ideas or working effectively with diverse teams (Dilmaghani, 2012; Dilmaghani and Dean, 2016). Other studies have indicated that religious individuals tend to have a higher degree of risk aversion (Miller and Hoffmann, 1995; Liu, 2010; Noussair et al., 2013), which may be negatively valued in the labor market (Heckman et al., 2006).

Many of these concerns are a moot point in our simple controlled task environment that involves individual decision-making. Therefore, we hypothesized that the positive aspects of religiosity on productivity (and honesty) should prevail and thus provide the clear contrast in productivity predictions between *ex-Prisoners* and *Religious* participants presented in H5.

5. RESULTS

5.1. Outcome variables of interest for the data analysis

HEADS reports and CHEATERS

The number of *HEADS* $\in$ [0,10] reported (out of 10 flips) is the simplest measure that can provide evidence of statistically likely cheating. This outcome measure is also used in our tests of moral disengagement—that is, do *HEADS* reports differ when completing the *Coin Flip* task after a 6-minutes versus 2-minutes real effort task, where the longer effort task implies a *lower* real wage rate for participation in the study. To identify a likely cheater, we also coded the binary variable *CHEATER* = 1 for those individuals reporting 8 or more HEADS flipped in the 10-flip task. Though this can occur by chance, the probability of *HEADS* $\geq$ 8 from 10 flips of a fair coin is approximately 5%. It is also the case that participants coded as *CHEATER* = 1 had significantly faster response times on the task ($p < .001$, 2-sample Z-test), which rejects a hypothesis that the same data generating process is at work in *CHEATER* = 1 versus *CHEATER* = 0 participants.

Fake Flippers (shirking—exploratory analysis)

Fake effort is related to shirking or impression management, and evidence suggests individuals proactively engage in such efforts.[23] We constructed a final binary, *Fake Flipper*, that has been used previously to identify participants who are statistically unlikely to be physically flipping a coin as requested (Dickinson and McEvoy, 2020; Dickinson, 2023). Because the *Coin Flip* task was administered in both online and in-lab formats in Dickinson and McEvoy (2020), we use their empirical distribution of response times from the in-lab version of the exact same task (administered using the exact same survey software) as the normative response time (RT) distribution one would observe if participants were actually flipping a coin as requested. They

[23] For an example, see https://www.cnn.com/2024/06/26/opinions/bossware-wells-fargo-mouse-jiggler-yang accessed July 19, 2024, which describes the use of mouse-jigglers to keep one's computer screen active.

scored those with RT < 45 seconds as *Fake Flippers* because this approximated the fastest 1% of the normative RT distribution. In other words, those with this fast of a task RT in our data were likely not actually flipping a coin. It is an important distinction that *Fake Flipper* = 1 need not imply a *Cheater* = 1. A fast RT may simply be a way to economize on time by cutting corners (from what was asked by the experimenter). Whether this proxies for a shirker, or impression management, it represents a behavior that could be costly in the workplace.

Productivity

In the real effort task, we use the number of characters input into the decoding box as a proxy for cumulative (unobservable) effort. For a comparable effort metric across participants assigned the 2-minute and 6-minute task, we then convert the outcome into *Productivity* = characters per minute decoded. Of course, this assessment of productivity does not examine the accuracy of the outcomes, but within the set of inaccurately decoded numeric strings, our methodology does not allow us to distinguish between inaccuracy by intention versus by inattention or incompetence.[24]

Reciprocity

Our last outcome measure sought to examine reciprocity (or, moral disengagement). The measure of this effect is captured by an interaction variable *Coin Flip Last * Long Task* where the dummy variable *Coin Flip Last* = 1 when the *Coin Flip* task is administrated after the effort task. If our hypothesis H3 is validated one should expect a positive coefficient for this interaction variable in predicting *HEADS,* as those completing the *Coin Flip* task last should cheat more if this followed a 6- versus 2-minutes effort task. Thus, a positive coefficient estimate on the *Coin Flip Last * Long Task* interaction would be evidence of negative reciprocity. That is, assignment of the lower real wage rate (i.e., longer effort task), compared to the higher real wage, may promote negative reciprocity, perhaps via moral disengagement, via dishonestly high*er HEADS* reports. Negative

[24] At least our informal review of the outcomes found very few instances (i.e., < 1%) of these data where individuals filled the text box with numeric (rather than letter-decoded) strings that are more clear instances of deliberate inaccuracy or severe task misunderstanding. In most all cases, the decoding effort shows evidence of good faith effort to accurately decode, with an occasional error or differences in trivial matters such as capitalization or omission of commas between letter strings (but still with spacing), or carriage returns at different points.

reciprocity could be the specific mechanism by which moral disengagement is manifest in our design, though we do not equate negative reciprocity to moral disengagement, in general.

Gift Exchange Outcomes (exploratory analysis)

While not a preregistered hypothesis, it is possible to examine the presence of a gift exchange in our data. Our experimental design makes clear to participants that one may either be assigned a 2- or 6-minutes real effort task for the exact same fixed study payment. In other words, those assigned the longer effort task by the experimenter are assigned a lower *wage rate*. As such, the participant's requested task effort may be seen as an act of reciprocity in response to the assigned task length (i.e., the *real* wage)—the gift exchange hypothesis (Akerlof, 1982). Though the classic gift exchange environment considers another study participant as the "employer" who has a payoff interest in the worker's effort, the environment here at least loosely represents one in which the participant may consider putting forth low effort in response to a low real wage rate assignment. We therefore conducted exploratory analysis to examine whether the *Long Task* treatment predicts effort/*Productivity*, and how personality trait or signal type may moderate this effect.

5.2. Study 1 findings: Dark versus Light personality traits

Table 2 provides summary measures regarding our variables of interest, though some hypotheses can only be examined in the regression analysis. We first test our preregistered hypotheses H1 concerning the *Coin Flip* task that measures dishonesty, and then follow with analysis of exploratory hypotheses. H1 is supported by simple tests of mean levels of *HEADS* reported, which showed that the average number of *HEADS* reported is statistically significantly higher than 5 out of 10 in the pooled data, as well as in both the *Dark* and *Light* subsamples (one-sample Z-tests, $p < .001$ in each case). Hypothesis 2 (H2) examines whether *Dark* types are more likely to be dishonest than *Light* types in the *Coin Flip* task.

[Table 2: about here]

Table 2 indicates that the average *HEADS* report is slightly higher for *Dark* (5.68) compared to *Light* (5.49) personality types, but a Mann-Whitney test shows that this difference is only marginally statistically significant (z=1.485; p=0.068; 1-tailed).[25] The proportion of those scored as *CHEATER* =1 participants (*HEADS* $\geq$ 8) is much higher among *Dark* compared to *Light* personality types, and the difference in proportions is highly statistically significant (Pearson X^2 = 7.1590; p = 0.0035; 1-tailed). Interestingly, if task response time (RT) is examined as another proxy for dishonesty, or as an exploratory measure of task shirking, 24% of *Dark* but only 17% of *Light* types were scored as *Fake Flipper* = 1 (Pearson X^2 = 5.6889; p=0.017; 2-tailed).

Tables 3 and 4 provide multivariate analysis on *HEADS* reports (Table 3) and complementary Probit estimations of the likelihood one is a *CHEATER* (Table 4). Table 3 shows the results of OLS estimates of the predictors of *HEADS* reports by *Dark* and *Light* trait clusters, with and without demographic and treatment controls. In both Tables 3 and 4 we also include in column (5) a model that uses the Big 5 personality characteristics (*Conscientiousness, Openness, Extraversion, Emotional Stability,* and *Agreeableness*) for comparison. Table 4 reports probit estimates of the *CHEATER* binary outcome variable. Figure 1 summarizes the results of the *HEADS* and *CHEATER* outcome analysis from Tables 3 and 4 along with complementary analysis using each individual trait (as opposed to clusters of traits) as the independent variable of interest (see Appendix Tables B2 and B3). Specifically, Figure 1 shows the coefficient point estimate for each trait's effect on the outcome measure along with the confidence interval on the estimated coefficient. The general findings are supportive of hypothesis H2 that *Dark* types are more likely to be dishonest than *Light* types. However, the results are most significant in the analysis of the *CHEATER* variable, where Table 4 shows a strong and robust effect of the *Light* personality cluster in decreasing the probability of being a *CHEATER* (p < .01 for the preregistered 1-tailed test). The *Dark* cluster is predicted to also increase the likelihood of being a *CHEATER,* although the statistical significance is lost when including demographic and treatment controls (model (2) of Table 4). We note also that, of the Big 5 characteristics, *Conscientiousness* positively predicts and *Agreeableness* negatively predicts the likelihood of being a *CHEATER*. Figure 1 and the Appendix

[25] We opted for 1-tailed tests as appropriate given our pre-registered hypotheses. For non-pre-registered and therefore exploratory results, we report 2-tailed tests throughout.

Tables B2 and B3 show some heterogeneity in the specific *Dark* and *Light* traits most responsible for associating traits with *CHEATER*. Notably, these are *Machiavellianism* and *Narcissism* that positively predict, and *Humanism* and *Kantianism* that negatively predict *CHEATER.* Additional sensitivity analysis in Appendix Tables B4 and B5 includes a sample selection correction, and the key findings survive this sensitivity analysis.[26,27] Finally, the *Females* ($p < .05$) and older participants (*Age* variable, $p < .01$) report lower *HEADS* outcomes in Table 3, but the effect of *Female* is not found in predicting *CHEATER* in Table 4. These demographic effects are generally consistent with previous literature.[28]

[Tables 3-4 and Figure 1: about here]

These findings are summarized in result 1.

Result 1. *a) There is general evidence of statistical cheating in the Coin Flip task. b) a relatively more Dark, compared to Light, personality type significantly increases HEADS reports and the likelihood of being classified a CHEATER.*

Alternatively, one might consider the variable *Fake Flipper* as a proxy for dishonesty, though more conservatively this is likely an indicator that the participant shirked the assigned task and did not flip a coin as requested. This analysis of *Fake Flipper* is exploratory (i.e., not pre-registered), and

[26] If one considers that our preregistered hypothesis does not apply to the related hypothesis of being scored a *CHEATER* =1 (given the preregistered hypothesis focused on *HEADS* outcomes), then our data still support a significant result of *Light* types being significantly less likely to be deemed a *CHEATER* using the 2-tailed test ($p = .011$). A similar result is found if using a combined *NetLight = Light Triad – Dark Triad* measure, which shows that *Netlight*, though not affecting *HEADS* reported, decreases the likelihood of being deemed a *CHEATER* ($p = .019$).

[27] The sample selection concern is addressed by estimated inverse-probability weight (IPW) corrected regressions (Appendix Table B4 and B5), where a first-stage probit regression was estimated on the entire set of those invited to participate in the study. That is, using observable characteristics available on participants in Dickinson (2023), we estimated the probability that an individual enrolled in our Study 1 as a function of those characteristics. The IPW approach then uses the inverse of that probability to give extra weight in the present analysis to those participants whose characteristics predicted a lower probability of enrollment in the study.

[28] Erat and Gneezy (2012) observe that men are more likely than women to lie for monetary gain in a cheap talk environment. Similarly, Houser et al. (2012) report that men are more likely than women to incorrectly report the result of a private coin flip. Concerning age effect, Glätzle-Rützler and Lergetporer, (2015) investigated how age influences the propensity to tell "black" and "white" lies in a sample of 383 children and teenagers aged 10/11 and 15/16 years. The authors find that a non-negligible fraction of subjects in both age cohorts exhibits lying-aversion and that the propensity to lie decreases significantly with age. In the context of an experiment on dishonesty, Fosgaard (2020) found that the older the participants are, the less they cheat.

we noted previously that *Dark* types were more likely to be *Fake Flippers* than *Light* types (Table 2: $p < .05$). Table 5 provides results of multivariate Probit estimations to further examine the impact of traits on *Fake Flipper.* Table 5 shows that more *Dark* personality positively predicts, and more *Light* personality negatively predicts, the likelihood of being a *Fake Flipper*. The complementary estimation results in Appendix B, Table B6, are summarized in Figure 2, which shows that *Narcissism* and *Psychopathy* predict an increased likelihood of being a *Fake Flipper* ($p < .05$), while *Humanism* and *Kantianism* predict a decreased likelihood of being a *Fake Flipper* ($p < .05$ and $p < .01$, respectively, for the two-tailed test results on the exploratory analysis). Extraversion also positively predicts the likelihood of *Fake Flipper* ($p < .01$). Further sensitivity analysis using the sample selection is found in Appendix B, Table B7, and the results are similar. These findings are summarized as follow:

Result 2. a) *Dark (Light) personality positively (negatively) predicts the likelihood of being a Fake Flipper; b) Narcissism and Psychopathy predict an increased likelihood of being a Fake Flipper; Humanism and Kantianism predict a decreased likelihood of being a Fake Flipper.*

[Tables 5: about here]

[FIGURE 2: about here]

Table 6 focuses on tests of reciprocity and personality-moderation Hypotheses H3 and H4 by showing the predictors of *HEADS* reports in the *Coin Flip* task for the full sample as well as for the separate subsamples of more *Dark* and *Light* types. Hypothesis H3 may also be examined by focusing on the same interaction term in Table 3, but we focus on Table 6 so that the H4 analysis of trait-moderation does not require the use of a double-interaction (e.g., *Coin Flip Last * Long Task * Dark trait*), which is difficult to interpret. In our Table 6, Hypothesis H4 is evaluated by comparison of the *Coin Flip Last * Long Task* interaction term across models (2) and (3). As is evident in Table 6, the Study 1 data do not support H3 or H4.

Result 3. *a) Assignment to a lower real pay rate for the effort task does not result in a subsequent increase in cheating or negative reciprocity towards the experimenter. b) We report no evidence that such reciprocity is moderated by Dark versus Light personality types.*

[Tables 6: about here]

Finally, the last preregistered Hypothesis 5 for Study 1 seeks to examine the impact of personality traits on /productivity in the real effort task, which we proxy with characters-per-minute decoded in the effort task. Table 2 reported that the average characters decoded per minute was 24.99 for *Dark* and 25.44 for *Light* personality types. A Wilcoxon Mann-Whitney test indicates that this difference is not statistically significant (z= 0.257; p=0.398; 1-tailed).

Table 7 shows results from regression analysis of the determinant of *Productivity* (analogous estimation results by individual traits are summarized in Figure 3—see also Appendix Table B8).

[Table 7 and Figure 3: about here]

Consistent with our non-parametric analysis, the coefficient associated with the *Dark* trait cluster is not significant in column (1). However, this variable becomes highly significant after controlling for demographics (see column 2, Table 7).[29] Here, the data show consistent support for Hypothesis 5. In Table 7 we see that, after controlling for demographics, the *Dark* trait personality cluster predicts significantly lower *Productivity*, while the *Light* traits cluster nominally predicts higher productivity, but with only marginal significance ($p < .10$). Appendix Table B8 complements the Table 7 findings by highlighting that all dark traits, except *Machiavellianism*, predict significantly lower productivity, while the positive traits impact on productivity fails to reach statistical significance (*Kantianism* and *Humanism* increase productivity marginally significantly at the $p <$ .10 level for the 1-tailed test of the preregistered hypothesis). Sensitivity analysis accounting for sample selection yields similar findings (see Appendix Table B9). Thus, these results are:

[29] It is important to remember that there were compositional differences across treatments in our sample, particularly in age and gender (see Table 1).

Result 4. *a) Dark types are significantly less productive than Light types; b) Narcissism, Psychopathy, and Sadism predict significantly lower productivity c) Kantianism and Humanism predict a marginally significant increase in productivity.*

5.3. Study 2 findings: Weak signals

We conducted similar analysis for Study 2, with the focus being on the binary indicator variable *ex-Prisoner*, which would identify effects of *ex-Prisoner* status relative to *Religious* participant status in our dataset. The analysis of Study 2 data is simpler than that of Study 1, because there are no individual sub-traits within each categorization to consider. We preregistered hypotheses that *ex-Prisoner* effects would be similar to *Dark* personality trait effects hypothesized in Study 1.

Hypothesis 1 is supported in the Study 2 data set, as we find that the average number of *HEADS* reported is statistically significantly higher than 5 out of 10 for both the *ex-Prisoner* and *Religious* subsamples of Study 2 data (see Table 2 summary statistics), as well as for the pooled sample (one-sample Z-tests, $p < .001$ in each case).

Table 8 summarizes the other findings from Study 2. Here, we present estimation models with the set of participant and task control measures as was done with Study 1. We do not report significant differences in *HEADS* reported or the likelihood of being scored a *CHEATER* between the two participant types. Thus, we fail to support H2 in the Study 2 data. Consistent with the exploratory finding from Study 1, *ex-Prisoner* participants had an estimated higher likelihood of being scored a *Fake Flipper* ($p = .018$), compared to *Religious* participants—the result from a non-parametric test of the Table proportions similarly conclude a significant difference in the proportion of participant from each group categorized as a *Fake Flipper* (Pearson $X^2 = 12.7454$, $p < 0.001$; 2-tailed test for the exploratory hypothesis). Secondly, and more surprisingly, we find that the *ex-Prisoner* participants were significantly *more* productive on the real effort task ($p = .001$), which is opposite our preregistered hypothesis. We do not have a clear explanation for this result. However, it should be interpreted with caution given the specific effort task we administered. One

might speculate that ex-prisoners are more accustomed to performing simple and repetitive tasks like those in the experiment compared to *Religious* participants. Results in the split sample estimations shown in Table 9, which examine hypotheses H3 and H4 related to cross-task reciprocity, reveals support for H4 in the *ex-Prisoner* subsample in column (1). The statistically significant and positive coefficient estimate on the interaction term *Coin Flip Last * Long Task* in the *ex-Prisoner* sample implies that *ex-Prisoner* status moderates a negative cross-task reciprocity.

[Table 8 and 9: about here]

Our overall Study 2 findings are summarized together as follow:

Result 5. *a) Ex-Prisoner participants had a higher likelihood of being scored a Fake Flipper compared to Religious participants. b) Compared to Religious participants, the ex-Prisoner participants were significantly more productive on the real effort task, contrary to our hypothesis. c) ex-Prisoner participants display a negative reciprocity effect*

To compare the estimated effects of ex-Prisoner and Dark personality more directly, we summarized the findings for both Study 1 and Study 2 visually in Figure 4 using coefficient plots—the upper panel shows main outcomes from Study 1, and the lower panel shows Study 2 results. Qualitatively, there are strong similarities between estimated effects of *Dark* types and *Ex-Prisoner*s (relative to *Lighter* and *Religious* types, respectively) for all outcome measures related to the *Coin Flip* task, though significance was greater for *Dark* types compared to *ex-Prisoners*. In contrast, *Dark* personality type predicts lower productivity on the real effort task, while *Ex-Prisoner* predicts higher productivity. While not shown in the coefficient plots, recall that there was some evidence for negative reciprocity in the *ex-Prisoner* sample, which was not found in any other subsample of participants from either study.

[FIGURE 4: about here]

5.4. Exploratory Analysis—Gift Exchange Behavior

To examine the (exploratory) gift exchange hypothesis in the data, we estimated models to predict standardized productivity as a function of task length, personality type, task order (*Coin Flip* before or after the *Effort* task), and demographic controls previously used. For this analysis, we omit the interaction between *Long Task* and *Coin Flip Last*, which was used in analysis of reciprocity that focused on *Coin Flip* outcomes. Here, a negative and significant coefficient estimate on the *Long Task* indicator variable would support the gift exchange hypothesis (i.e., per-minute productivity/effort is lower (higher) when assigned the lower (higher) real wage rate). We estimated the model of task *Productivity* separately for the Study 1 and Study 2 data sets as well as the separate subsample of participant types in each study. These results are shown in Table 10.

[Table 10: about here]

In Table 10, we focus our attention on the *Long Task* indicator variable coefficient estimates. Here, on for *Dark* types in Study 1 do we find evidence of a wage-effort gift exchange. That is, the negative coefficient estimate in column (2) implies that assignment to the longer effort task (i.e., the lower real wage) predicts lower effort—*Dark* types put forth more effort when assigned the shorter effort task. Of course, this exploratory finding assumes that characters-per-minute is a good proxy for task effort. It is also possible that one may explain this result as a type of negative reciprocity from *Dark* types when assigned the lower real wage rate. Thus, we have:

Exploratory Result 6: *Dark personality type predicts a marginally significant gift exchange in the effort task (or, alternatively, Dark types display significant negative reciprocity)*

6. DISCUSSION AND CONCLUSION

Firms face imperfect information in employee recruitment and, as a result, attempt to interpret signals of future productivity. This present study aimed to explore the validity of *Dark* versus *Light* personality types and *Religious* versus *ex-Prisoner* status, in predicting productivity, shirking, and honesty in incentived tasks. Study 1 results supported the notion that *Dark* relative to *Light* types are more likely to cheat and engage in fake effort in the *Coin Flip* task. They also put forth less effort in the real-effort task, although we found no evidence for negative cross-task reciprocity (i.e., moral disengagement) among *Dark* or *Light* types in Study 1. Further analysis revealed a

marginally significant gift exchange effort effect among *Dark* types compared to *Light* types. While the gift exchange result was exploratory, it suggests *Dark* types may subscribe to principles of reciprocity to a greater degree than *Light* types, or perhaps they are more sensitive to extrinsic incentives. We cannot rule out the possibility that estimated decreases in productivity when *Dark* types are assigned a low real wage rate represent a type of moral disengagement, because we are not able to discriminate pure negative reciprocity from an attempt to morally justify the behavior. Additional research should seek to examine the importance of reciprocity and its relationship with moral disengagement.

We also present some comparative findings with more traditional Big 5 personality characteristics in Study 1. Because Study 1 was designed to custom screen by *Dark* versus *Light* personality traits, this implies that our sample will not be representative of the personality types distributions typical in adult populations. Nevertheless, if one wishes to compare coefficient estimates across models that examine *Dark, Light,* or Big 5 characteristics effects on outcomes, we would first note that *Dark/Light* characteristics were measured on a 1-5 scale, whereas the Big 5 characteristics were measured along a 1-7 scale. Therefore, a 1-unit increase in the *Dark* cluster score, or a 1-unit increase in a specific trait (e.g., narcissism) implies a comparable movement along the Big 5 characteristic scale of 1.4 units (i.e., 20% movement along the 7-point scale). Even considering this, we would conclude that the magnitude of an effect for a similar marginal increase along the trait-scale is larger for the *Dark* and *Light* traits than it is for the Big 5 traits in a comparable regression that controls for gender and treatment effects.

Study 2 showed some different results among ostensibly comparable signal types (personality traits versus weak characteristics). *Religious* participants, compared to *ex-Prisoners*, were less productive in the effort task. Like *Dark* types, *ex-Prisoners* were more likely to be classified as shirking (i.e., *Fake Flipper*). However, *ex-Prisoners* did not report more *HEADS* and were not more likely to be classified a *CHEATER* in the Coin Flip task. On the contrary, Study 2 findings regarding cheating were more nuanced. First, we reported evidence of negative reciprocity, which could be interpreted as moral disengagement, in *Coin Flip* task outcomes with *ex-Prisoners*. That is, *ex-Prisoners* reported more *HEADS* in the coin flip task when it followed a 6-minutes compared

to a 2-minutes *Effort* task. Also, Study 2 results suggest that *Religious,* compared to *ex-Prisoner,* participants may not cheat more, in general, but when the *Coin Flip* task follows the effort task they reported significantly more *HEADS*. If, as some have suggested, self-control resources are required to behave honestly (Mead et al., 2009), then it may be more difficult to resist the temptation to cheat when this opportunity is at towards the end of a study rather than at the beginning. Though it is speculation, it may be that *Religious* participants may be more susceptible to this self-control depletion effect.

Some of the behavioral similarities and differences in seemingly related personality traits and acquired may suggest patterns of interest. For example, if *Dark* types are only as productive as *Light* types when offered a higher real wage rate, this may suggest that *Dark* types respond more to extrinsic motivation, whereas *Light* types may rely more heavily on intrinsic motivation in the workplace. Or, if *Ex-Prisoners* are more productive in menial effort tasks (compared to *Religious* types) then this may suggest trade-offs in terms of employee task assignments. A comparable tendency to engage more in fake effort or impression management may suggest similarities in how both *Dark* types and *ex-Prisoners* minimize the risk of cutting corners. Or, it may reflect how both types behaviorally respond to what may be viewed as an illegitimate or unnecessary task—some have considered such tasks as a vehicle with which one can wield supervisory abuse, and so such task assignment may be viewed as a type of punishment (Stein et al., 2020). Hopefully such speculation can help guide future research into such workplace behaviors.

Of course, we acknowledge our study has several limitations. Sample selection may be a concern in our data, particularly in our Study 2. Those who enrolled in this study may differ from representative *ex-Prisoner* or *Religious* individuals both in term of other observable or unobservable characteristics, which may affect external validity of our findings. Although this may be the case, we took precautions to reduce at least certain sources of sample bias in both our studies. First, the experiments were run online, which may reduce potential experimenter effects or social desirability bias. Secondly, participants were not aware that they were eligible for the study due to their *Dark/Light* personality trait or their *Religious/ex-Prisoner* characteristic. The Prolific platform merely presents studies on the participant dashboard for which the individual is

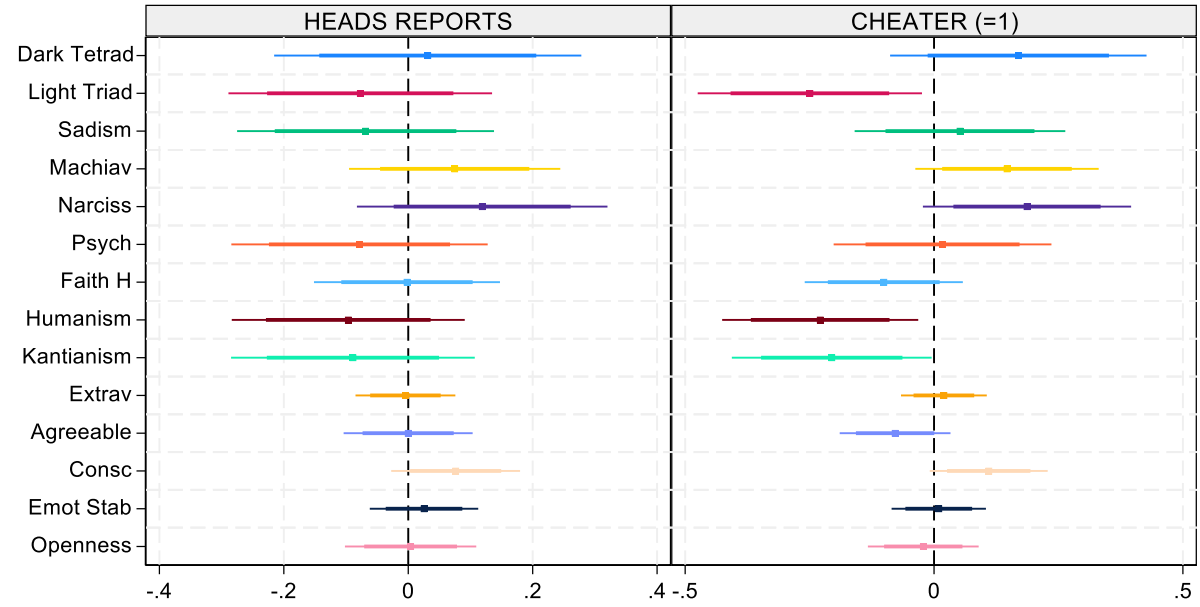
eligible. As such it is unlikely participants enrolled in the study with any subconscious bias to behave in accordance with any personality or characteristic stereotype. Regarding the *ex-Prisoner* sample in particular, we do not know in our data how long or how long ago the participant was in prison, nor do we know the type of offense for which they spent time in prison. Future studies can examine how these details further clarify the results we report.

Another limitation of this study relates to the tasks we administered. For instance, our *Coin Flip* task only identifies statistically likely cheating (though high *HEADS* reports have also been found to predict individually identifiable cheating in another task (Dickinson and Masclet, 2023)). The interpretation of our *Coin Flip* reciprocity finding in Study 2 is also complicated. For example, we interpreted the finding that *Religious* participants reported a main effect of more *HEADS* whenever the *Coin Flip* task was administered last as possible evidence of a self-control resources depletion effect. In other words, because one had already engaged in an effort task, self-control resources had been depleted, making one less able to resist the temptation or dishonest monetary gain (Baumeister and Exline, 1999; Wang et al, 2017). As such, an alternative interpretation of negative reciprocity effect among *ex-Prisoners* in the *Coin Flip* task is that self-control resources were even more depleted after a 6-minutes. Thus, *Ex-Prisoner* participants may experience self-control resource depletion more strongly than *Religious* participants. Additional research is needed to better understand the mechanism responsible for this finding. Regarding our *Effort* task, it may be considered too simple, or menial, to inform regarding several dimensions of workplace behavior. Individual repetitive tasks may not be as relevant certain workplace settings, or in environments where teamwork plays a major role. Those having spent time in prison may be more experienced in performing repetitive, menial, or perceived illegitimate tasks, which may explain the higher productivity and follow-up cheating relative to *Religious* participants.

Natural extensions of our research would be to investigate whether our findings hold in the context of alternative games and tasks, and also to explore the importance of individual-specific traits in more depth. While we deliberately focused on two tasks that produce simple outcome measures, they cannot capture the full breadth of tasks that contribute to an understanding of organizational citizenship and counterproductive workplace behaviors, both of which contribute in important

ways to corporate culture. Nevertheless, such building block tasks are often useful as a way to identify clear effects that improve our understanding of behaviors relevant in a wide variety of more complex environments. Our goal, ultimately, was to examine such building-block environments to better understand the relevance of personality traits and characteristics on key behaviors of general importance in organizations (i.e., honesty, effort, reciprocity). In the end, our results are at least generally supportive of efforts to solve the asymmetric information problem inherent in employee selection by using either personality traits or weak signals, because our results identify several areas in which these can help predict an increased likelihood of counterproductive workplace behaviors.

FIGURE 1: PREDICTORS OF HEADS AND CHEATING LIKELIHOOD



Notes: plot show the coefficient point estimates along with 99% (thin line) and 95% (fat lines) confidence interval.
Specifications included task and demographic controls

FIGURE 2: Fake Flipper (=1) by dark/light/Big 5 traits

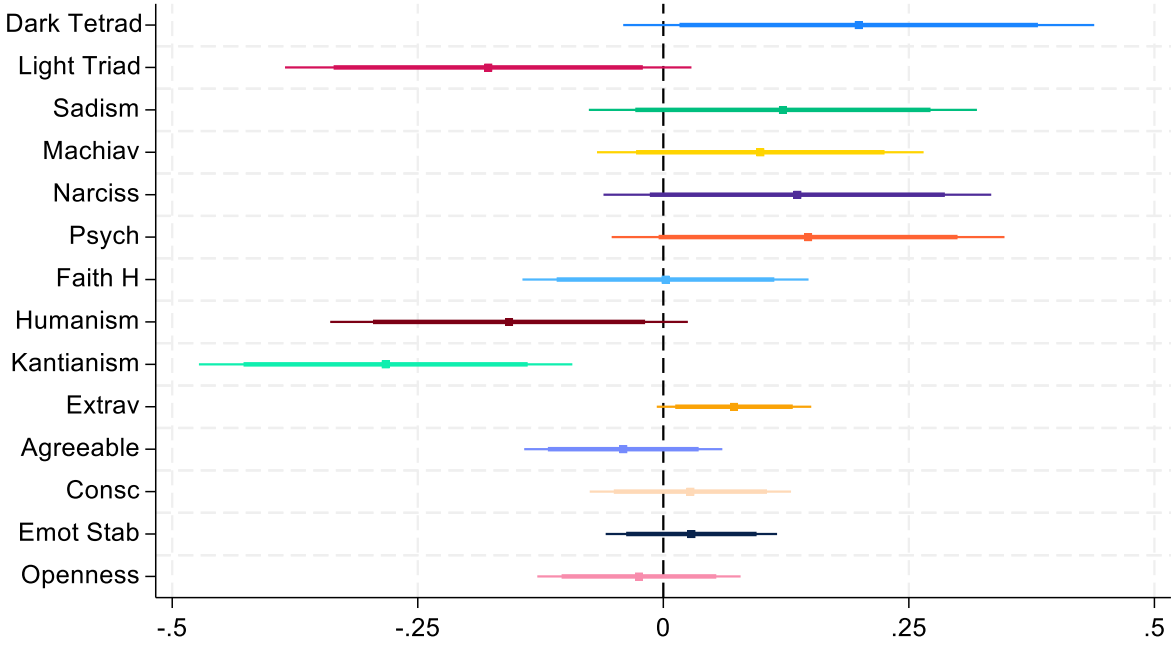


FIGURE 3: Task Productivity by dark/light/Big 5 traits

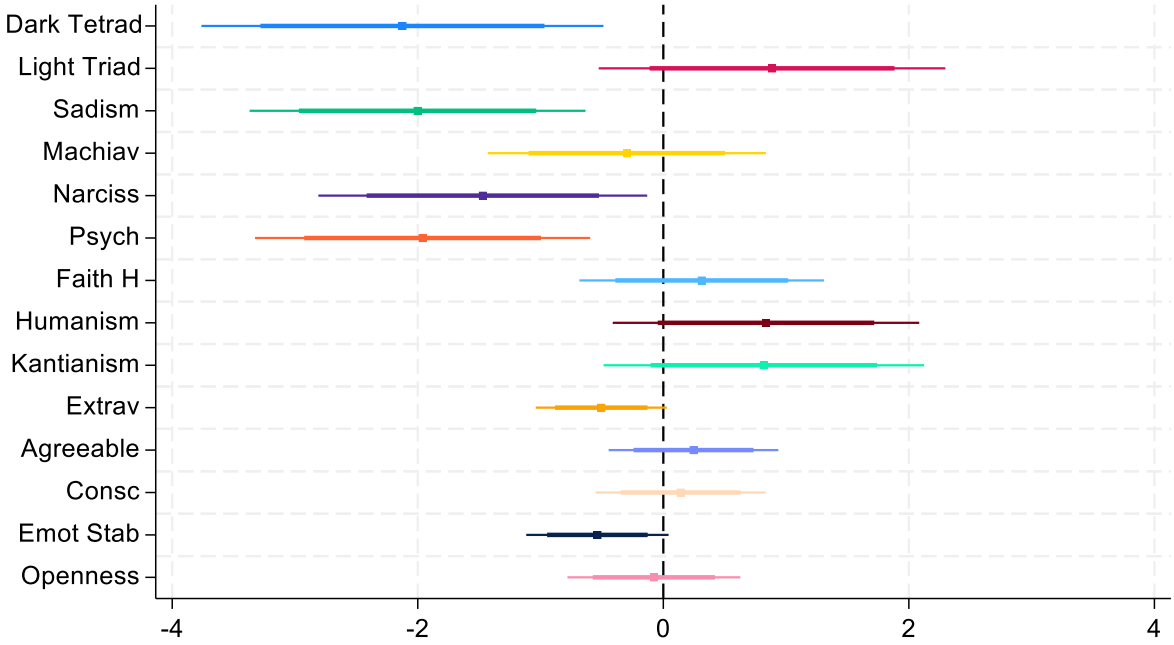


FIGURE 4 (Upper Panel): Impact of Dark traits score

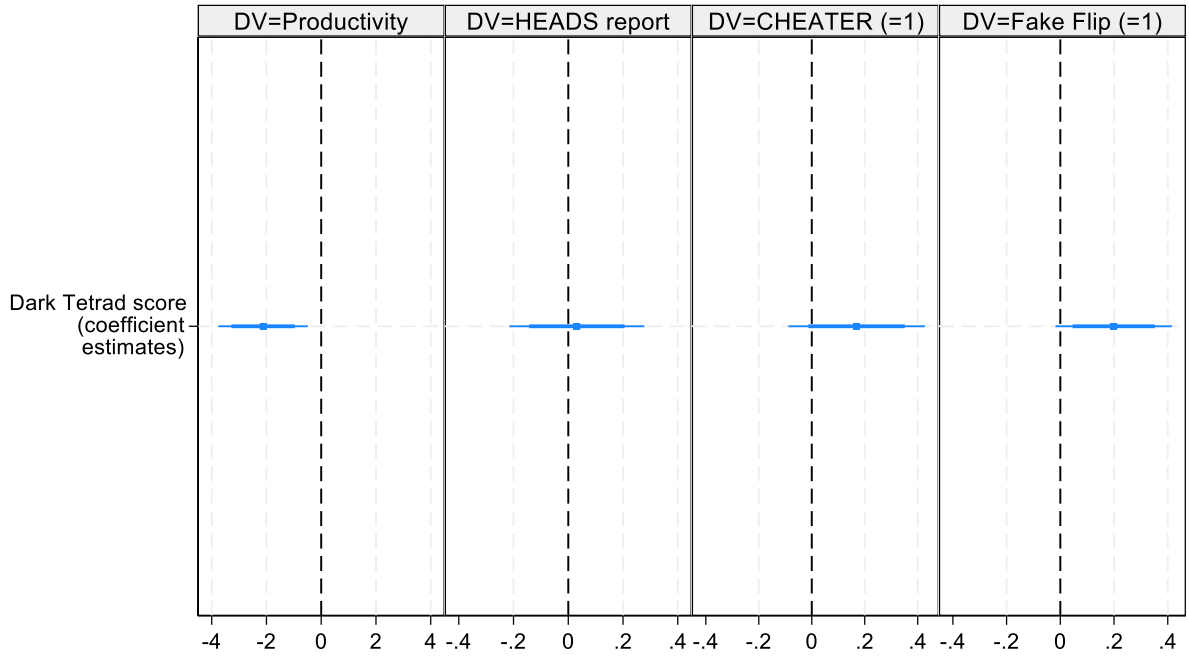
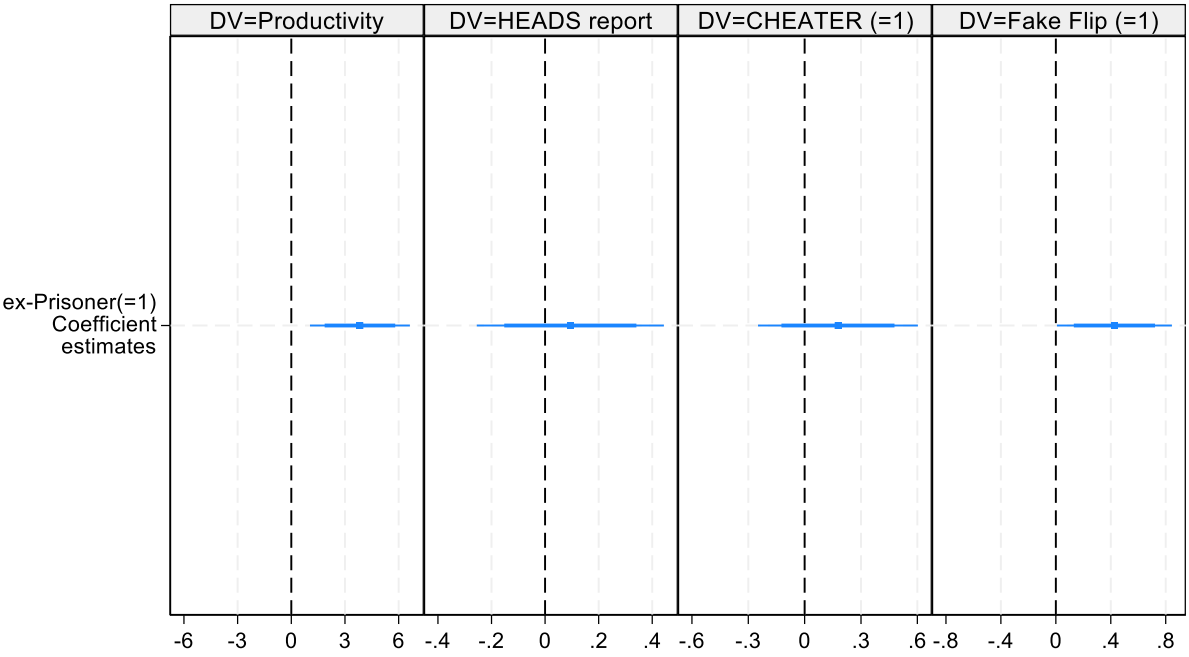


FIGURE 4 (Lower Panel): Impact of ex-Prisoner (=1)



Notes: Coefficient plots show the 95% (thick) and 99% (thin) confidence intervals for the 1-tailed test for preregistered hypotheses. Fake-Flip estimates are significant at $p < .05$ level for the appropriate 2-tailed test of the exploratory hypothesis. Results are similar in the Upper Panel for Study 1 if using the binary indicator for *Dark* versus *Light* subgroups, rather than the continuous *Dark Tetrad* measure.

TABLE 1: Demographics (age and sex) by treatment

Study 1	**Treatment** Pooled	**Treatment** Order AB 6 min effort	**Treatment** Order AB 2 min effort	**Treatment** Order BA 6 min effort	**Treatment** Order BA 2 min effort
Relatively Dark	N=399 (218 female) Age = 36.36 ± 12.28	n=95 (53 female) Age=36.15 ± 10.5	n=106 (55 female) Age=37.42 ± 12.78	n=104 (56 female) Age=36.96 ± 13.55	n=94 (54 female) Age=34.71 ± 11.87
Relatively Light	N=401 (306 female) Age=42.71 ± 14.35	n=86 (61 female) Age=42.10 ± 13.92	n=107 (92 female) Age=43.89 ± 13.52	n=117 (86 female) Age=41.50 ± 14.96	n=91 (67 female) Age=43.46 ± 14.95
Study 2	**Treatment** Pooled	**Treatment** Order AB 6 min effort	**Treatment** Order AB 2 min effort	**Treatment** Order BA 6 min effort	**Treatment** Order BA 2 min effort
Ex-Prisoner	N=297 (100 female) Age=39.42 ± 11.97	n=78 (29 female) Age=39.29 ± 10.67	n=69 (26 female) Age=38.23 ± 11.82	n=75 (21 female) Age=38.51 ± 13.06	n=75 (24 female) Age=41.56 ± 12.18
Religious	N=459 (290 female) Age=28.77 ± 9.55	n=118 (79 female) Age=28.95 ± 10.69	n=116 (72 female) Age=29.01 ± 9.19	n=112 (70 female) Age=29.22 ± 9.98	n=113 (69 female) Age=27.90 ± 8.20

Notes: Table shows mean values with standard deviations in parenthesis. For Study 1, *Relatively Dark* refers to those recruited from the lower quartile of the net-Light distribution, which captures the average dark trait measure subtracted from the average light trait measure from the database reported in Dickinson (2023) (*Relatively Light* are those from the upper quartile of the net-Light distribution). Treatment A refers to the effort task, Treatment B refers to the Coin Flip task. Also, "6 min effort" refers to being assigned the longer effort task, which implies a lower *real* wage for participation ("2 min effort" is the shorter effort task, or higher *real* wage assignment).

TABLE 2: Summary measure of outcomes of interest

	Coin flip task				Effort task
Study 1	**HEADS (Out of 10)**	**High Cheater (=1)**	**Coin Flip RT (seconds)**	**Fake Flipper (=1)**	**Productivity per minute**
Relatively Dark (n=399)	5.68 (1.73)	.12 (.33)	88.25 (80.37)	.24 (.43)	24.99 (10.35)
Relatively Light (n=401)	5.49 (1.51)	.07 (.25)	91.94 (56.09)	.17 (.38)	25.44 (11.81)
Study 2					
Ex-Prisoner (n=297)	5.42 (1.50)	.07 (.26)	96.10 (52.55)	.13 (.33)	24.29 (14.74)
Religious (n=524)	5.49 (1.54)	.07 (.26)	119.91 (54.41)	.05 (.23)	23.79 (10.49)

Notes: Table shows mean values with standard deviations in parenthesis

TABLE 3: HEADS reported by Personality Profile

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Dark Tetrad	Dark Tetrad	Light Triad	Light Triad	Big 5
Dark Tetrad	0.18*	0.03			
	(0.10)	(0.11)			
Light Triad			-0.16*	-0.08	
			(0.09)	(0.09)	
Conscientious					0.08
					(0.05)
Openness					-0.00
					(0.05)
Extraversion					-0.01
					(0.04)
Emotional Stability					0.01
					(0.04)
Agreeableness					-0.02
					(0.05)
Long Task		-0.06		-0.05	-0.07
		(0.16)		(0.16)	(0.16)
Coin Flip last (=1)		0.04		0.04	0.04
		(0.16)		(0.16)	(0.16)
Coin Flip Last * Long Task		0.10		0.10	0.11
		(0.23)		(0.23)	(0.23)
Age		-0.01**		-0.01**	-0.02**
		(0.00)		(0.00)	(0.00)
Female (=1)		-0.29*		-0.28*	-0.30*
		(0.13)		(0.12)	(0.13)
USA resident (=1)		0.12		0.12	0.12
		(0.12)		(0.12)	(0.12)
Constant	5.20**	6.17**	6.17**	6.51**	6.01**
	(0.22)	(0.38)	(0.34)	(0.38)	(0.37)
Observations	800	800	800	800	800
R-squared	0.00	0.02	0.00	0.02	0.03

Notes: ** p<0.01, * p<0.05 for the 1-tailed preregistered hypothesis of the personality effect (significance of other variables shown for 2-tailed tests). Standard errors in parentheses. Tobit estimations produce similar results, with 29 of 800 (3.6%) observations being censored.

TABLE 4: Likelihood of CHEATER (=1) by Personality Profile

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Dark Tetrad	Dark Tetrad	Light Triad	Light Triad	Big 5
Dark Tetrad	0.27**	0.17			
	(0.10)	(0.11)			
Light Triad			-0.29**	-0.25**	
			(0.09)	(0.10)	
Conscientious					0.15**
					(0.06)
Openness					-0.04
					(0.05)
Extraversion					0.03
					(0.04)
Emotional Stability					-0.01
					(0.05)
Agreeableness					-0.11*
					(0.05)
Long Task		0.17		0.18	0.16
		(0.18)		(0.18)	(0.19)
Coin Flip last (=1)		0.24		0.25	0.26
		(0.18)		(0.18)	(0.19)
Coin Flip Last * Long Task		-0.29		-0.31	-0.28
		(0.25)		(0.26)	(0.26)
Age		-0.01**		-0.01**	-0.02**
		(0.01)		(0.01)	(0.01)
Female (=1)		-0.18		-0.19	-0.24
		(0.13)		(0.13)	(0.14)
USA resident (=1)		0.05		0.05	0.10
		(0.13)		(0.13)	(0.13)
Constant	-1.92**	-1.21**	-0.23	0.10	-0.82
	(0.24)	(0.41)	(0.35)	(0.41)	(0.42)
Observations	800	800	800	800	800
Pseudo R-squared	0.0138	0.0352	0.0199	0.0439	0.0520

Notes: ** p<0.01, * p<0.05 for the 1-tailed preregistered hypothesis of the personality effect (significance of other variables shown for 2-tailed tests). Standard errors in parentheses.

TABLE 5: Fake Flipping by Personality Profile

Dependent Variable = *Fake Flipper* (=1)

Models are Probit estimations

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Dark Tetrad	Dark Tetrad	Light Triad	Light Triad	Big 5
Dark Tetrad	0.23**	0.20*			
	(0.08)	(0.09)			
Light Triad			-0.21**	-0.18*	
			(0.08)	(0.08)	
Conscientious					0.03
					(0.04)
Openness					-0.06
					(0.04)
Extraversion					0.08*
					(0.03)
Emotional Stability					0.02
					(0.04)
Agreeableness					-0.06
					(0.04)
Long Task		-0.13		-0.11	-0.12
		(0.15)		(0.15)	(0.15)
Coin Flip last (=1)		0.40**		0.40**	0.42**
		(0.14)		(0.14)	(0.14)
Coin Flip Last * Long Task		-0.14		-0.16	-0.14
		(0.21)		(0.21)	(0.21)
Age		-0.01*		-0.01**	-0.01**
		(0.00)		(0.00)	(0.00)
Female (=1)		0.05		0.02	-0.01
		(0.11)		(0.11)	(0.11)
USA resident (=1)		0.10		0.10	0.14
		(0.10)		(0.10)	(0.11)
Constant	-1.33**	-1.11**	-0.04	0.06	-0.48
	(0.20)	(0.35)	(0.29)	(0.33)	(0.34)
Observations	800	800	800	800	800
Pseudo R-squared	0.0094	0.0355	0.0088	0.0360	0.0420

Notes: ** p<0.01, * p<0.05 for the 2-tailed tests. Standard errors in parentheses.

TABLE 6: HEADS reports--Dark vs. Light subsamples

	(1)	(2)	(3)
VARIABLES	Full Sample	Dark subsample	Light subsample
Long Task (=1)	-0.06	-0.10	0.01
	(0.16)	(0.25)	(0.21)
Coin Flip last (=1)	0.04	-0.09	0.22
	(0.16)	(0.25)	(0.21)
Coin Flip Last * Long Task	0.10	0.27	-0.13
	(0.23)	(0.35)	(0.30)
Age	-0.01**	-0.02*	-0.01*
	(0.00)	(0.01)	(0.01)
Female (=1)	-0.30*	-0.16	-0.50**
	(0.12)	(0.18)	(0.18)
USA resident (=1)	0.12	0.13	0.07
	(0.12)	(0.18)	(0.15)
Constant	6.26**	6.38**	6.18**
	(0.23)	(0.36)	(0.32)
Observations	800	399	401
R-squared	0.02	0.02	0.03

Notes: ** p<0.01, * p<0.05 for the 1-tailed preregistered hypothesis of the personality effect (significance of other variables shown for 2-tailed tests). Standard errors in parentheses

TABLE 7: Productivity (effort) by Personality Profile

Dependent Variable = Characters decoded per minute

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Dark Tetrad	Dark Tetrad	Light Triad	Light Triad	Big 5
Dark Tetrad	-0.90	-2.12**			
	(0.67)	(0.70)			
Light Triad			0.39	0.89	
			(0.61)	(0.61)	
Conscientious					0.40
					(0.33)
Openness					0.15
					(0.32)
Extraversion					-0.45
					(0.24)
Emotional Stability					-0.69*
					(0.29)
Agreeableness					0.40
					(0.31)
Long Task		-0.89		-0.89	-1.12
		(1.07)		(1.07)	(1.08)
Coin Flip last (=1)		0.64		0.60	0.35
		(1.08)		(1.08)	(1.08)
Coin Flip Last * Long Task		-1.27		-1.20	-1.04
		(1.52)		(1.53)	(1.53)
Age		-0.19**		-0.17**	-0.16**
		(0.03)		(0.03)	(0.03)
Female (=1)		-0.83		-0.28	-0.67
		(0.84)		(0.82)	(0.85)
USA resident (=1)		2.94**		2.99**	2.97**
		(0.76)		(0.77)	(0.78)
Constant	27.19**	36.82**	23.76**	27.73**	30.45**
	(0.67)	(0.67)	(0.67)	(2.50)	(2.48)
Observations	800	800	800	800	800
R-squared	0.00	0.08	0.00	0.07	0.08

Notes: ** p<0.01, * p<0.05 for the 1-tailed preregistered hypothesis of the personality effect (significance of other variables shown for 2-tailed tests). Standard errors in parentheses. Tobit estimations produce similar results, but only 5 of 800 (< 1%) observations are censored.

TABLE 8: The effect of ex-Prisoner vs. Religious status

VARIABLES	(1) DV=Productivity	(2) DV=HEADS reported	(3) DV=CHEATER (=1)	(4) DV=FAKE FLIPPER (=1)
Ex-Prisoner (=1)	3.84**	0.09	0.18	0.43*
	(1.20)	(0.15)	(0.18)	(0.18)
Age	-0.23**	-0.01	-0.01	-0.00
	(0.04)	(0.01)	(0.01)	(0.01)
Female (=1)	0.52	0.09	0.07	-0.08
	(0.92)	(0.12)	(0.14)	(0.14)
Long Task (=1)	1.64	0.06	-0.02	-0.04
	(1.26)	(0.16)	(0.20)	(0.21)
Coin Flip last (=1)	1.20	0.11	-0.01	0.14
	(1.25)	(0.16)	(0.20)	(0.20)
Coin Flip Last * Long Task	-1.15	0.25	0.27	0.19
	(1.76)	(0.22)	(0.28)	(0.27)
USA resident (=1)	-1.74	-0.14	-0.17	0.16
	(1.35)	(0.17)	(0.21)	(0.18)
Constant	28.84**	5.49**	-1.30**	-1.55**
	(1.66)	(0.21)	(0.27)	(0.27)
Observations	756	756	756	756
R-squared	0.04	0.02		
Pseudo R-squared			0.01	0.04

Notes: ** p<0.01, * p<0.05 for the 2-tailed tests (1-tailed for preregistered hypotheses). N=756 observations. Standard errors in parentheses. Tobit estimations produce similar results for model columns (1) and (2), but only 10 (1.3%) Productivity outcomes and 13 (1.7%) HEADS reports are censored. Results are qualitatively similar if estimations use the full n=1033 Study 2 sample (where it is unobserved whether some of the *ex-Prisoners* may also be *Religious* (and vice-versa). In this case, however, the coefficient on the *Ex-Prisoner* indicator in the *Productivity* estimation is somewhat smaller and less statistically significant (β =2.42, p = .05) while the coefficient estimate on the *Ex-Prisoner* indicator in the *Fake Flipper* probit estimation is a bit larger and more precisely measured (β = .48, p = .001).

TABLE 9: HEADS reports--ex-Prisoner vs. Religious subsamples

VARIABLES	(1) Full Sample	(2) Ex-Prisoner subsample	(3) Religious subsample
Long Task (=1)	0.07	-0.16	0.20
	(0.16)	(0.25)	(0.20)
Coin Flip last (=1)	0.11	-0.38	0.42*
	(0.16)	(0.25)	(0.20)
Coin Flip Last * Long Task	0.25	0.79*	-0.08
	(0.22)	(0.35)	(0.29)
Age	-0.01	-0.01	-0.01
	(0.00)	(0.01)	(0.01)
Female (=1)	0.08	-0.09	0.22
	(0.11)	(0.19)	(0.15)
USA resident (=1)	-0.08	-0.16	0.03
	(0.15)	(0.18)	(0.55)
Constant	5.49**	5.90**	5.33**
	(0.21)	(0.37)	(0.27)
Observations	756	297	459
R-squared	0.02	0.03	0.03

Notes: ** p<0.01, * p<0.05 for the 1-tailed preregistered hypothesis of the personality effect (significance of other variables shown for 2-tailed tests). Standard errors in parentheses

TABLE 10: Gift Exchange by type/signal

Dependent variable = Productivity (per minute)

VARIABLES	(1) Study 1 full sample	(2) Study 1 Dark types	(3) Study 1 Light types	(4) Study 2 full sample	(5) Study 1 ex-Prisoner	(6) Study 1 Religious
Long Task (=1)	-1.45	-2.02*	-1.08	1.22	0.24	1.55
	(0.76)	(1.01)	(1.16)	(0.89)	(1.67)	(0.97)
Coin Flip last (=1)	-0.04	0.34	-0.36	0.56	1.94	-0.40
	(0.77)	(1.01)	(1.15)	(0.89)	(1.67)	(0.97)
Age	-0.17**	-0.17**	-0.19**	-0.18**	-0.28**	-0.15**
	(0.03)	(0.04)	(0.04)	(0.04)	(0.07)	(0.05)
Female (=1)	-0.04	-0.41	-0.35	-0.16	2.68	-1.11
	(0.80)	(1.03)	(1.35)	(0.90)	(1.78)	(1.01)
USA resident (=1)	3.04**	1.95	4.06**	0.39	-2.28	1.78
	(0.77)	(1.02)	(1.16)	(1.17)	(1.68)	(3.72)
Constant	30.92**	31.28**	32.35**	29.02**	34.50**	28.23**
	(1.49)	(2.06)	(2.38)	(1.60)	(3.37)	(1.76)
Observations	800	399	401	756	297	459
R-squared	0.07	0.07	0.08	0.03	0.08	0.03

Notes: ** p<0.01, * p<0.05, ^ p<.10 for the 2-tailed tests.

REFERENCES

Abbink, K., & Herrmann, B. (2011). The moral costs of nastiness. *Economic Inquiry*, *49*(2), 631-633.

Agan, A., & Starr, S. (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, *133*(1), 191-235.

Aghababaei, N., Mohammadtabar, S., & Saffarinia, M. (2014). Dirty dozen vs. the H factor: Comparison of the dark triad and honesty–humility in prosociality, religiosity, and happiness. *Personality and Individual Differences*, 67, 6–10.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, *97*(4), 543-569.

Albright, S., & Denq, F. (1996). Employer attitudes toward hiring ex-offenders. *The Prison Journal*, *76*(2), 118-137.

Ashton, M. C., Paunonen, S. V., Helmes, E., & Jackson, D. N. (1998). Kin altruism, reciprocal altruism, and the Big Five personality factors. *Evolution and Human Behavior*, *19*(4), 243-255.

Arrow, K., (1973). The theory of discrimination. Discrimination in labor markets 3. The Theory of Discrimination, S. 3-33 in: Orley Ashenfelter und Albert Rees (Hg.): Discrimination in Labor Markets

Austin, J., Irwin, J., & Kubrin, C. E. (2003). It's about time: America's imprisonment binge. *Punishment and Social Control*, 433-469.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*, *1986*(23-28).

Barsky, A. (2011). Investigating the effects of moral disengagement and participation on unethical work behavior. *Journal of Business Ethics*, 104(1), 59-75.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel Psychology*, *44*(1), 1-26.

Baumeister, R. F., & Juola Exline, J. (1999). Virtue, personality, and social relations: Self-control as the moral muscle. *Journal of Personality*, *67*(6), 1165-1194.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*(2), 169-217.

Bicchieri C. (2006). The Grammar of Society: The nature and Dynamics of Social Norms. *Cambridge University press*, Cambridge, MA.

Birkeland, S., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2014). An experimental study of prosocial motivation among criminals. *Experimental Economics*, *17*, 501-511.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317-335.

Boshier, R., & Johnson, D. (1974). Does conviction affect employment opportunities. *British Journal of Criminology*, *14*, 264.

Buikhuisen, Wouter, and Fokke PH Dijksterhuis. "Delinquency and stigmatisation." *Brit. J. Criminology* 11 (1971): 185.

Byford, K. U. (1995). The quest for the honest worker: A proposal for regulation of integrity testing. *SMU Law Review*, *49*, 329-373.

Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society interface*, *18*(175), 20200880.

Carpenter, J., P. Matthews, and J. Schirm (2010), "Tournaments and Office Politics: Evidence from a real effort experiment," *American Economic Review*, 100(1), 504-517.

Charness, G., Masclet, D., & Villeval, M. C. (2014). The dark side of competition for status. *Management Science*, *60*(1), 38-55.

Christian, J. S., & Ellis, A. P. (2014). The crucial role of turnover intentions in transforming moral disengagement into deviant behavior at work. *Journal of Business Ethics*, *119*, 193-208.

Coffin, B. (2003). Breaking the silence on white collar crime. *Risk Management*, *50*(9), 8-9.

Cohn, A., Maréchal, M. A., & Noll, T. (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies*, *82*(4), 1289-1308.

Conroy, S. J., & Emerson, T. L. (2004). Business ethics and religion: Religiosity as a predictor of ethical awareness among students. *Journal of Business Ethics*, *50*, 383-396.

Decker, S. H., Ortiz, N., Spohn, C., & Hedberg, E. (2015). Criminal stigma, race, and ethnicity: The consequences of imprisonment for employment. *Journal of Criminal Justice*, *43*(2), 108-121.

Detert, J. R., Treviño, L. K., & Sweitzer, V. L. (2008). Moral disengagement in ethical decision making: A study of antecedents and outcomes. *Journal of Applied Psychology*, 93(2), 374-391.

Dickinson, D. L., (2023). Dark versus light personality types and moral choice. *IZA Discussion Paper, No. 16338*.

Dickinson, D. L., & Masclet, D. (2023). Unethical decision making and sleep restriction: Experimental evidence. *Games and Economic Behavior*, *141*, 484-502.

Dickinson, D. L., & McEvoy, D. M. (2021). Further from the truth: The impact of moving from in-person to online settings on dishonest behavior. *Journal of Behavioral and Experimental Economics*, *90*, 101649.

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, *99*(1), 197-199.

Egan, V., Hughes, N., & Palmer, E. J. (2015). Moral disengagement, the dark triad, and unethical consumer attitudes. *Personality and Individual Differences*, *76*, 123-128.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723-733.

Fernández-del-Río, E., Ramos-Villagrasa, P. J., & Barrada, J. R. (2020). Bad guys perform better? The incremental predictive validity of the Dark Tetrad over Big Five and Honesty-Humility. *Personality and Individual Differences*, *154*, 109700.

Forsyth, D. R., Banks, G. C., & McDaniel, M. A. (2012). A meta-analysis of the Dark Triad and work behavior: a social exchange perspective. *Journal of Applied Psychology*, *97*(3), 557.

Fosgaard, T. R. (2020). Students cheat more: Comparing the dishonesty of a student sample and a representative sample in the laboratory. *The Scandinavian Journal of Economics*, *122*(1), 257-279.

Glätzle-Rützler, D., & Lergetporer, P. (2015). Lying and age: An experimental study. *Journal of Economic Psychology*, *46*, 12-25.

Goldberg, E. (2023, March 5). The Value Of Looking Beyond The Resume. *New York Times*, 6(L). https://link.gale.com/apps/doc/B739568442/BIC?u=boon41269&sid=bookmark-BIC&xid=c68c23e9

Gøtzsche-Astrup, O., Overgaard, B., & Lindekilde, L. (2022). Vulnerable and dominant: Bright and dark side personality traits and values of individuals in organized crime in Denmark. *Scandinavian Journal of Psychology*, *63*(5), 536-544.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Stanford University Press.

Hanson, G. A. (1991). To catch a thief: The legal and policy implications of honesty testing in the workplace. *Law & Inequality.*, *9(3)*, 497.

Harbring, C., B. Irlenbusch, M. Kräckel, and R. Selten (2007), "Sabotage in Asymmetric Contests - An Experimental Analysis," *International Journal of the Economics and Business*, 14, 201-223.

Harris, L. L., Jackson, S. B., Owens, J., & Seybert, N. (2021). Recruiting dark personalities for earnings management. *Journal of Business Ethics*, 1-26.

Harrison, A., Summers, J., & Mennecke, B. (2018). The effects of the dark triad on unethical behavior. *Journal of Business Ethics*, *153*, 53-77.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Heller, M. (2005). Court ruling that employer's integrity test violated ADA could open door to litigation. *Workforce Management*, *84*(9), 74-77.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial–organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology*, *1*(3), 272-290.

Houser, D., Vetter, S., & Winter, J. (2012). Fairness and cheating. *European Economic Review*, *56*(8), 1645-1655.

Huesmann, L. R., Dubow, E. F., & Boxer, P. (2011). The effect of religious participation on aggression over one's lifetime and across generations. In J. P. Forgas, A. W. Kruglanski, & K. D. Williams (Eds.), The psychology of social conflict and aggression (pp. 301–322). Sydney, Australia: Sydney University Press.

James, S., Kavanagh, P. S., Jonason, P. K., Chonody, J. M., & Scrutton, H. E. (2014). The Dark Triad, schadenfreude, and sensational interests: Dark personalities, dark emotions, and dark behaviors. *Personality and Individual Differences*, *68*, 211-216

Kämmerle, M., Unterrainer, H. F., Dahmen-Wassenberg, P., Fink, A., & Kapfhammer, H. P.(2014). Dimensions of religious/spiritual well-being and the dark triad of personality. *Psychopathology*, 47(5), 297–302.

Kaufman, S. B., Yaden, D. B., Hyde, E., & Tsukayama, E. (2019). The light vs. dark triad of personality: Contrasting two very different profiles of human nature. *Frontiers in Psychology, 10*, 467.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608-638.

Koch, A., Nafziger, J., and Nielsen, H. S. (2015). Behavioral economics of education. *Journal of Economic Behavior & Organization*, 115:3–17.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association*, *11*(3), 495-524.

Lazear, E. P. (1989). Pay equality and industrial politics. *Journal of Political Economy*, *97*(3), 561-580.

Lazear, E. P., & Gibbs, M. (2014). *Personnel economics in practice*. John Wiley & Sons.

LeBreton, J. M., Shiverdecker, L. K., & Grimaldi, E. M. (2018). The dark triad and workplace behavior. *Annual Review of Organizational Psychology and Organizational Behavior*, *5*, 387-414.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives*, *21*(2), 153-174.

List, J. A., C. D. Bailey, P. J.Euzent, and T. L.Martin (2001). "Academic economists behaving badly? A survey on three areas of unethical behavior," *Economic Inquiry*, 39(1), 162–170.

Łowicki, P., & Zajenkowski, M. (2017). No empathy for people nor for God: The relationship between the Dark Triad, religiosity and empathy. *Personality and Individual Differences*, *115*, 169-173.

March, E., & Marrington, J. Z. (2021). Antisocial and prosocial online behaviour: Exploring the roles of the dark and light triads. *Current Psychology*, 1-4.

Marcus, B., & Schuler, H. (2004). Antecedents of counterproductive behavior at work: a general perspective. *Journal of Applied Psychology*, *89*(4), 647.

Masclet, D., & Dickinson, D.L. (2024). Incorporating conditional morality into economic decisions. *Theory and Decision* (forthcoming).

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology*, *45*(3), 594-597.

Mercado, B. K., Giordano, C., & Dilchert, S. (2017). A meta-analytic investigation of cyberloafing. *Career Development International*, *22*(5), 546-564.

Miao, Y., Wang, J., Shen, R., & Wang, D. (2023). Effects of Big Five, HEXACO, and Dark Triad on Counterproductive Work Behaviors: A Meta-Analysis. *International Journal of Mental Helath Promotion*, *25*(3), 357-374.

Moore, C. (2015). Moral disengagement. *Current Opinion in Psychology*, *6*, 199-204.

Moore, C., Detert, J. R., Treviño, L. K., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology*, 65(1), 1-48.

Murphy, K. R. (1993). *Honesty in the workplace*. Thomson Brooks/Cole Publishing Co.

Newman, A., Le, H., North-Samardzic, A., & Cohen, M. (2020). Moral disengagement at work: A review and research agenda. *Journal of Business Ethics*, *167*, 535-570.

Nguyen, N., Pascart, S., & Borteyrou, X. (2021). The dark triad personality traits and work behaviors: A person-centered approach. *Personality and Individual Differences*, *170*, 110432.

Oda, R., & Matsumoto-Oda, A. (2022). HEXACO, Dark Triad and altruism in daily life. *Personality and Individual Differences*, *185*, 111303.

Olsen, K. J., & Stekelberg, J. (2016). CEO narcissism and corporate tax sheltering. *The Journal of the American Taxation Association*, *38*(1), 1-22.

O'Reilly, C. A., & Hall, N. (2021). Grandiose narcissists and decision making: Impulsive, overconfident, and skeptical of experts–but seldom in doubt. *Personality and Individual Differences*, *168*, 110280.

Palmer, J. A., & Tackett, S. (2018). An examination of the Dark Triad constructs with regard to prosocial behavior. *Acta Psychopathologica*, *4*(5), 10-4172.

Paul Annie Murphy (2004) "You are what you Score" Free Press

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, *62*(4), 659-661.

Piazza, J. (2012). "If you love me keep my commandments": Religiosity increases preference for rule-based moral arguments. *International Journal for the Psychology of Religion*, 22, 285–302.

Piazza, J., & Landy, J. F. (2013). "Lean not on your own understanding": Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision making*, *8*(6), 639-661.

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*, *5*(3), 334-342.

Pletzer, J. L., Oostrom, J. K., & de Vries, R. E. (2021). HEXACO personality and organizational citizenship behavior: A domain-and facet-level meta-analysis. *Human Performance*, *34*(2), 126-147.

Preston, I., and S. Szymanski (2003), "Cheating in Contests," *Oxford Review of Economic Policy*, 19(4), 612-624.

Rijsenbilt, A., & Commandeur, H. (2013). Narcissus enters the courtroom: CEO narcissism and fraud. *Journal of Business Ethics*, *117*, 413-429.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support?. *Human Resource Management Review*, *16*(2), 155-180

Saroglou, V. (2002). Religion and the five factors of personality: A meta-analytic review. *Personality and Individual Differences*, 32(1), 15–25

Schwartz, R. D., & Skolnick, J. H. (1962). Two studies of legal stigma. *Social Problems*, *10*(2), 133-142.

Schwieren, C., & Weichselbaumer, D. (2010). Does competition enhance performance or cheating? A laboratory experiment. *Journal of Economic Psychology*, *31*(3), 241-253.

Sevi, B., Urganci, B., & Sakman, E. (2020). Who cheats? An examination of light and dark personality traits as predictors of infidelity. *Personality and Individual Differences*, *164*, 110126

Shleifer, A. (2004), "Does Competition Destroy Ethical Behavior?," *American Economic Review Papers and Proceedings*, 94(2), 414-418

Spence, M. (1973). Job Market Signaling, *The Quarterly Journal of Economics*, vol. 87(3), pp. 355-374.

Steers, R., & Rhodes, S. (1984). *Knowledge and speculation about absenteeism*. In P. Goodman & R. Atkin (Eds.), Absenteeism, vol. 1: 229–275. San Francisco: Jossey-Bass

Stein, M., Vincent-Höper, S., Schümann, M., & Gregersen, S. (2020). Beyond mistreatment at the relationship level: Abusive supervision and illegitimate tasks. *International Journal of Environmental Research and Public Health*, *17*(8), 2722.

Sternberg, R. J. & Wagner, R. K. (1993). The geocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science*, vol. 2(1), pp. 1-4.

Sutter, M., Kocher, M. G., Glatzle-Rutzler, D., and Trautmann, S. T. (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review*, 103(1):510–31.Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Science*, 7, 320–324

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*(4), 703-742.

Van Scotter, J. R., & Roglio, K. D. D. (2020). CEO bright and dark personality: Effects on ethical misconduct. *Journal of Business Ethics*, *164*, 451-475.

Vize, C. E., Collison, K. L., Miller, J. D., & Lynam, D. R. (2020). Using item-level analyses to better understand the consequences of partialing procedures: An example using the Dark Triad. *Journal of Personality*, *88*(4), 719-734.

Vitell, S. J. (2009). The role of religiosity in business and consumer ethics: A review of the literature. *Journal of Business Ethics*, *90*, 155-167.

Waldfogel, J. (1994). The effect of criminal conviction on income and the trust" reposed in the workmen". *Journal of Human Resources*, 62-81.

Wang, Y., Wang, G., Chen, Q., & Li, L. (2017). Depletion, moral identity, and unethical behavior: Why people behave unethically after self-control exertion. *Consciousness and Cognition*, *56*, 188-198.

Williams, W. & Ceci, S. (1997). "How'm I Doing?" Problems with Student Ratings of Instructors and Courses. *Change*, vol. 29(5), 12. Retrieved Friday, April 13, 2007 from the ERIC database.

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review*, *12*(2), 250-264.

Zettler, I. (2022). A glimpse into prosociality at work. *Current Opinion in Psychology*, *44*, 140-145.

APPENDIX A: Theoretical framework

We present a theoretical framework designed to generate testable implications regarding the relationship between personality traits (or weak signals) and moral choices in the coin flip task, as well as performance in the real effort task. Our model is inspired by Masclet and Dickinson (2019). This model is based on the intuitive idea that individuals care about their own material payoffs but they have also some moral concerns such that any deviation from one's moral obligation may induce disutility. Another important idea behind this model is that morality is weak as individuals may revise their moral target upward (downward) by observing others' behaviors and how they are treated by others. Specifically, moral obligation is a combination of both an autonomous moral imperative component and a social influence component. We argue here that both components are influenced by personality traits. Our model attempts to provide testable hypotheses regarding how personality traits affect moral obligation and therefore individual decisions. We do not claim this model to be the only framework that may be useful in this regard, but we argue that a framework for decision making with moral concerns may help identify key pathways through which personality traits may affect both honesty and effort choices.

A.1. General framework: weak moral motivation and personality traits

One practical extension of Masclet and Dickinson (2019)'s model may be that individuals may be endowed with certain "dark" or "light" personality traits and that such traits may influence their moral target. Let's assume that each individual *i* is characterized by different personality traits represented in a vector $\boldsymbol{P_i}$. This vector has *n* components so, $\boldsymbol{P_i} = (P_{i1}, \dots, P_{in})$. These personality traits may include both dark (such as narcissism, psychopathy, Machiavellianism, sadism, etc.) and light personality components (for instance Faith in humanity, Humanism, Kantianism, etc.).[30] For simplicity, let's consider that personality traits are either considered as dark traits P_{id} or light traits P_{iL}. Consequently, we can summarize the vector $\boldsymbol{P_i}$, as follow: $\boldsymbol{P_i} = (P_{id}, P_{iL})$.

[30] For simplicity, we implicitly assume here that some weak signals such as the fact of being religious or an ex-prisoner may be explained by various personality traits included in vector $\boldsymbol{P}$.

Following Masclet and Dickinson (2019) we can represent the individual i' utility function as follows:

$$U(a_i) = b(a_i) - c(a_i) - v_i(a_i - \hat{a}_i(\boldsymbol{P_i})) \qquad (1)$$

where a_i is an action that generates both benefits, *b*, and costs, *c*. Both benefits and costs are twice continuously differentiable: $b' > 0$, $c' > 0$, $b'' \leq 0$, $c'' \geq 0$. The morality component of the utility function is captured by $v(a_i - \hat{a}_i(\boldsymbol{P_i}))$, which subtracts from utility for actions that deviate from one's moral imperative, $\hat{a}_i(\boldsymbol{P_i})$, in either direction—$v'_a > 0$ if $a > \hat{a}$, $v'_a < 0$ if $a < \hat{a}$, and $v'_a = 0$ if $a_i = \hat{a}_i$. Also, it is assumed that marginal disutility increases at an increasing rate as one's action gets further from the moral obligation such that $v''_{aa} > 0$. The moral imperative may refer, for instance, to a moral obligation to behave honestly in one's life or to the intrinsic motivation related to an innate sense of duty to exert high work effort in the context of workplace including self-esteem, interest and pride in one's work, an innate sense of duty to honor contractual obligations (Deci, 1975; Baron, 1988; Kreps, 1997; James, 2005; Ellingsen and Johanesson, 2008; Kuhnen and Tymula, 2012). Deviations of one's action from this moral obligation generate disutility (e.g., Nyborg, 2000; Brekke et al., 2003; Figuieres et al., 2013). [31]

As in Masclet and Dickinson (2019), we assume that the moral obligation component, $\hat{a}_i$, includes both an autonomous moral component denoted K_i (Laffont, 1975; Harsanyi, 1980) and a conditional moral component that is a function of social influence and fairness considerations denoted $F_i(a_j)$, where a_j is the action of others $j \neq i$. The moral obligation function can therefore be written as: : $\hat{a}_\iota(\boldsymbol{P_i}) = \hat{a}_\iota\left(K_i(\boldsymbol{P_i}), F_i(a_j, \boldsymbol{P_i})\right), j \neq i$ with $\hat{a}'_K \geq 0$ (one's moral obligation is weakly increasing in one's autonomous moral component) and $\hat{a}'_F \geq 0$ (one's moral obligation is weakly increasing in the perceived morality of others' behavior).

[31] Alternatively, we may also conjecture that personality traits may simply distort the morality function v_i by giving more or less weight to moral considerations depending on whether the traits are dark or light. For instance, one may reasonably argue that those with Kantian personality traits place a very high importance on morality. On the contrary, those who have very dark traits can be completely deprived from moral considerations, so that v_i=0. We can thus rewrite the utility function as follows: $U(a_i) = b(a_i) - c(a_i) - \boldsymbol{P_i} v_i(a_i - \hat{a}_i)$
However, this specification does not consider the fact that personality traits can influence not only the autonomous moral component but also social influence. For this reason, a specification where personality traits directly influence moral obligation seems more appropriate here.

Let's consider first the autonomous component $K_i(\boldsymbol{P_i})$. This autonomous component corresponds to the moral ideal in the absence of any social influence. It can be evolutionarily anchored or result from previous interactions (such as with our parents during the process of education and the transmission of norms that have been internalized). This autonomous component of the moral obligation function, $K_i(\boldsymbol{P_i})$, satisfies the intuitive property:

ASSUMPTION 1: $\frac{\partial K_i}{\partial P_{iL}} > 0$; $\frac{\partial K_i}{\partial P_{iD}} < 0$ iif $K_i \geq 0$ and $\frac{\partial K_i}{\partial P_{iL}} < 0$; $\frac{\partial K_i}{\partial P_{iD}} > 0$ iif $K_i < 0$

This assumption is quite obvious: those with light personality traits have a higher (lower) autonomous (a)moral component, so that : $\frac{\partial K_i}{\partial P_{iL}} > 0$ if $K_i \geq 0$ $\left(\text{and } \frac{\partial K_i}{\partial P_{iL}} < 0 \text{ if } K_i < 0\right)$. On the contrary, those who have dark personality traits have a lower (lower) autonomous (a)moral component such that $\frac{\partial K_i}{\partial P_{iD}} < 0$ if $K_i \geq 0$ $\left(\text{and } \frac{\partial K_i}{\partial P_{iD}} > 0 \text{ if } K_i < 0\right)$.

We now turn to the second component of the moral target, namely the social influence component $F_i\left(a_j, \boldsymbol{P_i}\right)$. This social influence component is defined as follows:

$$F_i\left(a_j, \boldsymbol{P_i}\right) = \lambda(\boldsymbol{P_i})\left[\frac{\left(a_j - a_j^{ref}\right)}{\left(a_j^{max} - a_j^{min}\right)}\left[a_i^{max} - a_i^{min}\right] + a_i^{min}\right] \qquad (2)$$

Where a_j is individual *j*'s action in the set that contains all possible actions from minimal to maximal, $a_j \in A_j = [a_j^{min}, a_j^{max}]$. a_j^{ref} is the reference point for considering whether the action of the other player *j* is fair or unfair. If player *i* feels he is treated badly by player *j*, (because $a_j < a_j^{ref}$), then he would revise downward his intrinsic moral ideal obligation. Alternatively, player *i* would positively reciprocate a fair action (because $a_j \geq a_j^{ref}$) by upwardly revising his moral motivation. The parameter $\lambda(\boldsymbol{P_i})$ captures the weight associated to the social influence function to illustrates the fact that the answer to social influence (fairness) can be stronger or weaker depending on personality traits. Precisely, the two following intuitive assumptions about the role of personality traits on $F_i\left(a_j, \boldsymbol{P_i}\right)$ are as follows:

ASSUMPTION 2: $\frac{\partial F_i(a_j, \boldsymbol{P_i})}{\partial P_{iD}} > 0$ iif $a_j < a_j^{ref}$; $\frac{\partial F_i(a_j, \boldsymbol{P_i})}{\partial P_{iD}} = 0$ iif $a_j \geq a_j^{ref}$

ASSUMPTION 3: $\frac{\partial F_i(a_j,\boldsymbol{P_i})}{\partial P_{iL}} < 0$ iif $a_j < a_j^{ref}$; $\frac{\partial F_i(a_j,\boldsymbol{P_i})}{\partial P_{iL}} > 0$ iif $a_j \geq a_j^{ref}$

The assumption 2 indicates that those with dark personality traits are more prone to reciprocate negatively when they feel they are badly treated (i.e. because $a_j < a_j^{ref}$). In other words, people with high dark personality traits are more likely to overreact negatively to others' actions, to exhibit more aggressive and retaliatory behaviors when they feel challenged or disrespected. Previous studies have shown that this may be the case that individuals with dark personality traits—such as narcissism, psychopathy, and Machiavellianism—are more prone to reciprocate negatively in social interactions (e.g. Kaufman et al. 2019). The fact that we assume that $\frac{\partial F_i(a_j,\boldsymbol{P_i})}{\partial P_{iD}} = 0$ if $a_j \geq a_j^{ref}$ suggests that those with darker personality traits are not supposed to overreact positively when they feel they are fairly treated. Indeed, previous studies using self - reports have found that none of the dark triad measure are significantly correlated with reciprocal altruism (i.e. Oda et al. 2022).[32]

The third assumption means that those with a light personality are more inclined to reciprocate positively $\frac{\partial F_i(a_j,\boldsymbol{P_i})}{\partial P_{iL}} > 0$ iif $a_j \geq a_j^{ref}$ and less inclined to reciprocate negatively $\frac{\partial F_i(a_j,\boldsymbol{P_i})}{\partial P_{iL}} < 0$ iif $a_j < a_j^{ref}$. Indeed, previous studies have found that positive traits correlate with positive reciprocity (Ashton et al. 1998) but negatively correlated with "reactive" aggression (e.g. Kaufman et al., 2019).

An illustration of the weak moral motivation function may be the following (e.g. Figueres et al. 2013):

$\hat{a}_{ij} = (1-\theta_i)K_i(\boldsymbol{P_i}) + \theta_i F_i(w_{ij}, \boldsymbol{P_i})$.

The weight θ_i may be interpreted as the conditionality of *i*'s moral motivation. If $\theta_i = 0$, individual *i* has strong unconditional moral motivation, and such an individual never deviates from his ideal moral intrinsic target K_i no matter the observed action of others.

[32] Other studies found mixed evidence showing that Machivellianism and Psychopathy are negatively correlated with positive reciprocity while Narcissism is positively correlated with reciprocal altruism (Palmer and Tackett, 2018).

A.2. Weak moral motivation and personality traits in the effort task

Let's now describe how our theoretical framework produces testable implications regarding personality traits and effort level in our effort task. Remind that our real effort task consisted in decoding a series of 5-digit numbers into 5-letter blocks and participants were paid a fixed wage *w*. It was common knowledge to participants that they may have to either complete the task for 2 or 6 minutes, which was chosen by the experimenter, which indirectly affects the wage rate per minute.

In the absence of any moral concerns, and under the assumption of perfect rationality and selfishness, standard theoretical predictions are straightforward. Given that effort is costly and that players are paid a fixed wage, the equilibrium of this game corresponds to the lowest effort possible, i.e. zero effort. However, a large body of experimental evidence including field and lab experiments have shown that, despite the absence of any penalty for shirking, workers do not hesitate to exert positive effort under a fixed wage scheme (e.g. Falk and Ichino, 2006; Mas and Moretti, 2009; Dohmen and Falk, 2010; Armentier and Boly, 2011; Greiner et al., 2011; Kuhnen and Tymula, 2012; Charness et al., 2014). These findings suggest that individuals derive some utility from exerting effort (or choosing above-minimal effort). This is consistent with a hypothesis of intrinsic motivation that can capture self-esteem, interest and pride in one's work, an innate sense of duty to honor contractual obligations (Baron, 1988; Kreps, 1997; James, 2005; Ellingsen and Johanesson, 2008), or a sense of fulfillment (Deci, 1975; Kuhnen and Tymula, 2012).Our model can account for such intrinsic motivation.

Precisely, let us consider the worker's payoff function in the effort task with moral concerns:

$$U_i(a_i, w_{ij}) = w_{ij} - c(a_i) - v_i(a_i - \widehat{a_i}) \qquad \text{with } \widehat{a_i} = \widehat{a_i}\left(K_i(\boldsymbol{P_i}), F_i(w_{ij}, \boldsymbol{P_i})\right) \tag{3}$$

Where $a_i \in [0, \overline{a}]$ is the effort level and w_{ij} correspond to the fixed wage that firm *j* (here the experimenter) offers worker *i*. Let's assume that w_{ij} can be either low or high such that $w_{ji} \in [\underline{w}, \overline{w}]$, which corresponds either to a low wage rate per minute when participants have to complete the task for 6 minutes or a high wage rate when the task must be performed during 2 minutes only. $c(a_i)$ is worker *i*'s cost of effort function (where *c*' > 0 and *c*'' > 0). To keep matters simple, we

can specify the cost function by considering a simple disutility of effort function: $c(a_i) = \delta a_i^2$. As noted previously, $v_i(a_i - \widehat{a_i})$ is the "moral obligation" function that generates disutility when effort differs from one's moral ideal, $\hat{a}_i$, and this moral ideal is a function of both unconditional moral motivation, K_i, and a social influence component that depends on the wage w_{ij}received by the firm j : $\hat{a}_i = \widehat{a_i}\left(K_i(\boldsymbol{P_i}), F_i\left(w_{ij}, \boldsymbol{P_i}\right)\right)$. For simplicity, we assume that *i*'s moral motivation is captured by a quadratic function such that $v_i(a_{ij} - \hat{a}_i) = (a_{ij} - \hat{a}_i)^2$

The autonomous component $K_i(\boldsymbol{P_i})$ refers to ones' intrinsic moral motivation (e.g. Deci, 1975; Kuhnen and Tymula, 2012) that depends on ones' personality traits such that $\frac{\partial K_i}{\partial P_{iL}} > 0$ and $\frac{\partial K_i}{\partial P_{iD}} < 0$.

The social influence component $F_i\left(w_{ij}, \boldsymbol{P_i}\right)$ can be interpreted as worker *i*'s perception regarding firm *j*'s fairness, where a high wage is perceived as an act of kindness, such that $\frac{\partial \hat{a}_i}{\partial w_{ij}} > 0$. The social influence component corresponds to :

$$F_i\left(w_{ij}, \boldsymbol{P_i}\right) = \lambda(\boldsymbol{P_i})\left[\frac{\left(w_{ij} - w_{ij}^{ref}\right)}{\left(w_j^{max} - w_j^{min}\right)}\left[a_i^{max} - a_i^{min}\right] + a_i^{min}\right]$$

where w_{ij}^{ref} is the reference wage which can consist of either the average wage on the labor market or the reservation wage, i.e. the minimum acceptable wage for worker *i*. If the received wage is below this reference wage ($w_{ij} < w_{ij}^{ref}$), the social influence component becomes negative and worker *i* revises downward his intrinsic moral ideal obligation as he feels he is treated badly by the firm. In the opposite, if the received wage is above the reference point ($w_{ij} \geq w_{ij}^{ref}$), the social influence function becomes positive, which automatically leads worker *i* to upwardly revise his moral motivation. Based on assumptions 2 and 3, we assume that the the reciprocal reaction depends on the dark or light nature of the personality traits in vector $\boldsymbol{P_i}$.

Each worker *i* chooses her effort level a_i to maximize:

$$max_{a_i}\ w_{ij} - c(a_i) - v_i(a_i - \widehat{a_i}), \quad \text{with } \widehat{a_i} = \widehat{a_i}\left(K_i(\boldsymbol{P_i}), F_i\left(w_{ij}, \boldsymbol{P_i}\right)\right) \qquad (4)$$

We get the following first order condition:

$$\text{FOC: } \frac{\partial U}{\partial a_i}: -c_i^{'}(a_i) - v_i^{'}(a_i - \hat{a}_i) = 0 \qquad \text{with } \hat{a}_\iota = \hat{a}_\iota\left(K_i(\boldsymbol{P_i}), F_i\left(w_{ij}, \boldsymbol{P_i}\right)\right) \tag{5}$$

This FOC can be solved to obtain Nash equilibrium effort level $a_i^* = a^*(w_{ij}, \boldsymbol{P_i})$, where both the received wage and personality traits influence effort. The following identity holds when substituting optimal effort and the moral obligation function back into (5):

$$-c'_a\left(a^*(w_{ij}, \boldsymbol{P_i})\right) - v_a{}'\left(a^*(w_{ij}, \boldsymbol{P_i}) - \hat{a}_\iota\left(K_i(\boldsymbol{P_i}), F_i\left(w_{ij}, \boldsymbol{P_i}\right)\right)\right) \equiv 0 \tag{6}$$

By differentiating both sides of this identity with respect to the wage, w_{ij}, we get:

$$\left(-c''_{aa} - v''_{aa}\right)\frac{\partial a^*}{\partial w_{ij}} - v''_{a\hat{a}}\frac{\partial \hat{a}}{\partial w_{ij}} = 0 \tag{7}$$

From equation (7) we can then get the following comparative static result:

$$\frac{\partial a_i^*}{\partial w_{ij}} = \frac{v''_{a\hat{a}}\left(\frac{\partial \hat{a}}{\partial w}\right)}{-c''_{aa} - v''_{aa}} > 0 \tag{8}$$

Since both $c(\cdot)$ and $v(\cdot)$ are convex functions, the denominator is unambiguously negative, and $\frac{\partial \hat{a}}{\partial w}$ is positive by assumption. This implies that the necessary condition for the existence of a positive wage effort reciprocity is that $v''_{a\hat{a}} < 0$. This condition is true by assumption, but recall that the interpretation of this condition is that a marginal increase in the moral obligation (resulting from increased wage by employer) raises the marginal gain to increased work effort on the part of the worker in term of a marginal reduction in moral disutility.

Now let's differentiate both sides of the identity (7) with respect to the personality traits, P_{iD} and P_{iL}, respectively. We get:

$$\left(-c''_{aa} - v''_{aa}\right)\frac{\partial a^*}{\partial P_{iD}} - v''_{a\hat{a}}\frac{\partial \hat{a}}{\partial P_{iD}} = 0 \tag{9a}$$

$$(-c''_{aa} - v''_{aa})\,\frac{\partial a^*}{\partial P_{iL}} - v''_{a\hat{a}}\,\frac{\partial \hat{a}}{\partial P_{iL}} = 0 \qquad (9b)$$

From equations (9a) and (9b) we can then get the following comparative static results:

$$\frac{\partial a_i^*}{\partial P_{iD}} = \frac{v''_{a\hat{a}}\left[\frac{\partial \hat{a}}{\partial P_{iD}}\right]}{-c''_{aa}-v''_{aa}} < 0 \qquad (10a) \quad \text{and} \quad \frac{\partial a_i^*}{\partial P_{iL}} = \frac{v''_{a\hat{a}}\left[\frac{\partial \hat{a}}{\partial P_{iL}}\right]}{-c''_{aa}-v''_{aa}} > 0 \qquad (10b)$$

As mentioned above, both $c(\cdot)$ and $v(\cdot)$ are convex functions, so the denominator is unambiguously negative. $v''_{a\hat{a}}$ is also negative. The sign of $\frac{\partial \hat{a}}{\partial P_{iD}}$ in eq. (10a) is negative since by assumption dark traits negatively affect one's autonomous moral component $\frac{\partial K_i(\boldsymbol{P_i})}{\partial P_{iD}} < 0$ and reinforce the *negative* social influence component $\frac{\partial F_i\left(w_{ij},\boldsymbol{P_i}\right)}{\partial P_{iD}}>0$ when $w_{ij} < w_{ij}^{ref}$ or leave it unchanged $\frac{\partial F_i(w_{ij},\boldsymbol{P_i})}{\partial P_{iD}}=0$ when $w_{ij} \geq w_{ij}^{ref}$.

The sign of $\frac{\partial \hat{a}}{\partial P_{iL}}$ is positive as light personality traits affect positively the autonomous component $\frac{\partial K_i(\boldsymbol{P_i})}{\partial P_{iL}} > 0$ and reinforce positive reciprocity $\frac{\partial F_i\left(w_{ij},\boldsymbol{P_i}\right)}{\partial P_{iL}} > 0$ if $w_{ij} \geq w_{ij}^{ref}$ while negative reciprocity is attenuated for light personalities $\frac{\partial F_i\left(w_{ij},\boldsymbol{P_i}\right)}{\partial P_{iL}} < 0$ if $w_{ij} < w_{ij}^{ref}$.

Altogether eq. (10a) and (10b) imply that dark (light) personality traits negatively (positively) affect the moral obligation $\hat{a}$, which translates into a lower (higher) effort level in the effort task.

To summarize, our model indicates that intrinsic moral motivation coupled with reciprocity may explain why workers choose above minimal effort under a flat wage scheme and how personality traits may affect effort levels and may shape the reciprocity function.

For illustration, let's replace the cost function and the moral concern function by their values into equation (5) and we get the following FOC:

$$-2\delta a_i - 2a_i + 2\left((1-\theta_i)K_i(\boldsymbol{P_i}) + \theta_i F_i\left(w_{ij},\boldsymbol{P_i}\right)\right) = 0 \qquad (11)$$

Thus, we have:

$$a_i^*(w_{ij}) = \frac{\left((1-\theta_i)K_i(\boldsymbol{P_i})+\theta_i F_i\left(w_{ij},\boldsymbol{P_i}\right)\right)}{(1+\delta)} \quad (12)$$

It is apparent from (12) that without moral concerns (*K* and *F*) effort should be zero. Interestingly we can also see that personality traits may also affect either positively or negatively effort level depending on the type dark or light of such traits.

A.3. Weak moral motivation and personality traits in the coin flip task

Let's now describe how our theoretical framework provides testable implications regarding how personality traits affect decisions in the coin flip task. Let's rewrite eq. (3) as follows:

$$U(a) = b(a) - v_i(a_i - \hat{a}_i(\boldsymbol{P_i})) \quad (13)$$

Where a_i describes dishonesty in the coin flip task (i.e. reporting HEADS when the coin flip says TAILS) that generates material benefits $b(a)$. In absence of material cost or penalties for cheating, the only costs incurred by a cheater are moral costs reflected by our moral function $v_i(a_i, \hat{a}_i(\boldsymbol{P_i}))$.

Since we are now dealing with the negative domain (i.e. "the dark side of human nature"), $K_i(\boldsymbol{P_i})$ now should be interpreted differently compared to the previous context of the effort task. Precisely, a positive value of $K_i(\boldsymbol{P_i})$ means that player *i* derives non-material benefits from cheating. In contrast, if $K_i(\boldsymbol{P_i}) < 0$ player *i* incurs disutility from cheating. This implies that assumptions 1-3 should be considered with the opposite sign to what was presented earlier.

It should be remembered that in our experimental design, the coin flip game was played either at the beginning or after the effort game. This may be of importance when considering the

effect of our moral function on cheating decisions. Indeed, when the coin flip task is played first the moral target sums to $\hat{a}_\iota = \hat{a}_i\big(K_i(\boldsymbol{P_i})\big)$ as there are no social influence. In contrast, when the flip coin is played after the effort task, the moral target may include a social influence component due to the fact that the decision in the flip coin may potentially be influenced by how the individual feels he was well or badly treated in the effort task by receiving either a low or a high wage. Consequently, in this latter case the moral target becomes $\hat{a}_\iota = \hat{a}_\iota\left(K_i(\boldsymbol{P_i}), F_i\big(w_{ij}, \boldsymbol{P_i}\big)\right)$.

Let's consider first the situation where the coin flip task is played first and assume also that the moral motivation is a quadratic function given by $v_i(a_i - \hat{a}_i) = (a_i - \hat{a}_i)^2 = \big(a_i - K_i(\boldsymbol{P_i})\big)^2$

Maximization of (13) yields the following FOC with respect to action levels:

$$b - 2a_i + 2[K_i(\boldsymbol{P_i})] = 0 \tag{14}$$

Now we have the optimal level of cheating activity a_i^* as:

$$a_i^* = \frac{b+2[K_i(\boldsymbol{P_i})]}{2} \tag{15}$$

We can easily see the following from (15) that cheating decreases with unconditional ethical concerns (i.e., if $K_i(\boldsymbol{P_i}) < 0$) and it increases (decreases) with dark (light) personality traits.

Let's consider now the situation where the coin flip task is played after the effort task. The moral obligation is now given by $\hat{a}_{ij} = (1 - \theta_i)K_i(\boldsymbol{P_i}) + \theta_i F_i\big(w_{ij}, \boldsymbol{P_i}\big)$. We get the following moral function :

$v_i(a_i - \hat{a}_i) = (a_i - \hat{a}_i)^2 = \left(a_i - \left((1 - \theta_i)K_i(\boldsymbol{P_i}) + \theta_i F_i\big(w_{ij}, \boldsymbol{P_i}\big)\right)\right)^2$

We egt the following FOC:

$b - 2a_i + 2\big[(1 - \theta_i)K_i(\boldsymbol{P_i}) + \theta_i F_i\big(w_{ij}, \boldsymbol{P_i}\big)\big] = 0$

Withe the optimal level of cheating activity a_i^* as:

$$a_i^{**} = \frac{b+2\left[(1-\theta_i)K_i(\boldsymbol{P_i})+\theta_i F_i\left(w_{ij},\boldsymbol{P_i}\right)\right]}{2} \quad (16)$$

It can be easily seen from comparison between eq. (16) and (15) that effort level $a_i^{**} \geq a_i^*$ for light personalities due to the additional social influence that may reinforce reciprocity and thus the moral target. In sharp contrast, for similar reasons one may expect lower effort for dark personalities ($a_i^{**} \leq a_i^*$).

To summarize, our theoretical framework provides the following testable hypotheses for our effort and coin flip tasks:

Hypothesis 1: a) *The number of HEADS reported will be greater than 5 (i.e., we hypothesize statistical evidence of cheating in the Coin Flip* t*ask. b) Dark, relative to Light, personality traits will report more HEADS in the Coin Flip task.*

Hypothesis 2: *Dark, compared to Light, personality traits, will put forth less effort in the real effort task.*

Hypothesis 3: *Those assigned to lower wage rate (i.e. a longer real effort) will exert lower effort, particularly if they have Dark personality traits.*

Hypothesis 4 : *a)Those assigned to lower wage rate (i.e. a longer real effort task) will cheat more on the coin flip task, when the coin flip task is administered after the effort task. b) This effect will be stronger for those relatively higher in Dark personality traits.*

Appendix A References

Armentier O, & Boly A. 2011. A controlled field experiment on corruption. *European Economic Review,* 55: 1072-1082.

Ashton, M. C., Paunonen, S. V., Helmes, E., & Jackson, D. N. (1998). Kin altruism, reciprocal altruism, and the Big Five personality factors. *Evolution and Human Behavior*, *19*(4), 243-255.

Baron J. 1988. The employment relation as a social relation. *Journal of the Japanese and International Economy*, 2(4): 492- 525.

Brekke KS, Kverndokk S, & Nyborg K. 2003. An economic model of moral motivation. *Journal of Public Economics,* 87: 1967–1983.

Charness G, Masclet D, & Villeval MC. 2014. The dark side of competition for status. *Management Science,* 60(1): 38-55.
Deci EL. 1975. *Intrinsic Motivation.* New York: Plenum Publishing Corp.
Dohmen T. & Falk A. 2010. Performance pay and multi-dimensional sorting productivity, Preferences and Gender. *American Economic Review,* 101(2): 556-590.
Ellingsen T. & Johannesson M. 2008. Pride and prejudice: The human side of incentive theory, *American Economic Review*, 98: 990-1008.
Falk A, & Ichino A. 2006. Clean evidence on peer pressure. *Journal of Labor Economics,* 24(1): 39-57.
Figuieres C, Masclet D, & Willinger M. 2013. Weak moral motivation leads to the decline of voluntary contributions. *Journal of Public Economic Theory*, 15(5): 745-772.
Greiner B, Ockenfels A, & Werner P. 2011. Wage transparency and performance: A real effort experiment. *Economic Letters,* 111: 236-238.
Harsanyi J. 1980. Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory and Decision*, 12: 115-133.
James H. 2005. Why did you do that? An economic examination of the effect of extrinsic compensation on intrinsic motivation and performance. *Journal of Economic Psychology*, 26(4): 549-566.
Kaufman, S. B., Yaden, D. B., Hyde, E., & Tsukayama, E. (2019). The light vs. dark triad of personality: Contrasting two very different profiles of human nature. *Frontiers in psychology*, *10*, 467.
Kuhnen C, & Tymula A. 2012. Feedback, self-esteem and performance in organizations. *Management Science*, 58: 94-113.
Kreps D. 1997. Intrinsic motivation and extrinsic incentives. *American Economic Review*, 87(2): 359-364.
Laffont JJ. 1975. Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. *Economica*, 42(168): 430-437.
Mas A, & Moretti E. 2009. Peers at work. *American Economic Review*, 99(1): 112-45.
Masclet, D. & Dickinson, D. L. (2019). Incorporating conditional morality into economic decisions. *IZA Discussion Paper No.* 12872
Nyborg K. 2000. Homo economicus and homo politicus: Interpretation and aggregation of environmental values. *Journal of Economic Behavior and Organization*, 42: 305–322.
Oda, R., & Matsumoto-Oda, A. (2022). HEXACO, Dark Triad and altruism in daily life. *Personality and Individual Differences*, *185*, 111303.
Palmer, J. A., & Tackett, S. (2018). An examination of the Dark Triad constructs with regard to prosocial behavior. *Acta Psychopathologica*, *4*(5), 10-4172.

APPENDIX B: Additional Tables (supporting main manuscript Figures and sensitivity analysis)

TABLE A1: Correlation Matrix--Dark and Light Clusters and Big 5

Light Tetrad	Conscientiousness	Openness	Extraversion	Emotional Stability	Agreeableness
1.00					
0.27	1.00				
0.16	0.12	1.00			
0.17	0.10	0.30	1.00		
0.19	0.42	0.17	0.21	1.00	
0.65	0.30	0.15	0.09	0.29	1.00

TABLE A2: *HEADS* reported by Dark/Light/Big5 Traits--OLS

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narc	Psych	FaithH	Human	Kant	Extrav	Agreeable	Conscien	EmotS	Open
Dark tetrad	0.03													
	(0.11)													
Light Triad		-0.08												
		(0.09)												
Sadism			-0.07											
			(0.09)											
Machiav				0.07										
				(0.07)										
Narcissism					0.12									
					(0.09)									
Psychopathy						-0.08								
						(0.09)								
FaithHum							-0.00							
							(0.06)							
Humanism								-0.10						
								(0.08)						
Kantianism									-0.09					
									(0.08)					
Extraversion										-0.00				
										(0.03)				
Agreeable											-0.000			
											(0.044)			
Conscient												0.076		
												(0.044)		
EmotStab													0.025	
													(0.037)	
Openness														0.004
														(0.045)
R-squared	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.023	0.027	0.024	0.023

Notes: *p<.05, **p<.01. Coefficient estimates shown (standard errors in parenthesis). Model column titles highlight independent variable used for model, which are OLS regressions that include controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=800 observations in all models.

TABLE A3: *CHEATER* (=1) by Dark/Light/Big5 Traits (Probit models)

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narciss	Psych	Faith H	Human	Kant	Extrav	Agree	Conscien	EmotS	Openness
Dark tetrad	0.17													
	(0.11)													
Light Triad		-0.25**												
		(0.10)												
Sadism			0.05											
			(0.09)											
Machiav				0.15*										
				(0.08)										
Narcissism					0.19*									
					(0.09)									
Psychop						0.02								
						(0.09)								
FaithHum							-0.10							
							(0.07)							
Humanism								-0.23**						
								(0.08)						
Kantianism									-0.21**					
									(0.09)					
Extraversion										0.02				
										(0.04)				
Agreeable											-0.078			
											(0.048)			
Conscient												0.110*		
												(0.051)		
EmotStab													0.010	
													(0.041)	
Openness														-0.022
														(0.048)
Pseudo R-squared	0.035	0.044	0.031	0.037	0.039	0.031	0.035	0.045	0.042	0.031	0.03	0.04	0.031	0.031

Notes: *p<.05, **p<.01. Coefficient estimates shown (standard errors in parenthesis). Model column titles highlight independent variable used for model, which are Non-linear Probit estimations regressions that include controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=800 observations in all models.

TABLE A4: *HEADS* reported by Dark/Light/Big5 Traits--IPW Correction

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narc	Psych	FaithH	Human	Kant	Extrav	Agreeable	Conscien	EmotS	Open
Dark tetrad	0.07													
	(0.11)													
Light Triad		-0.10												
		(0.10)												
Sadism			-0.04											
			(0.09)											
Machiav				0.11										
				(0.08)										
Narcissism					0.12									
					(0.09)									
Psychopathy						-0.05								
						(0.09)								
FaithH							-0.01							
							(0.07)							
Humanism								-0.12						
								(0.08)						
Kantianism									-0.12					
									(0.09)					
Extraversion										-0.01				
										(0.03)				
Agreeable											-0.037			
											(0.047)			
Conscient												0.071		
												(0.044)		
EmotStab													0.033	
													(0.039)	
Openness														0.026
														(0.051)
R-squared	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.020	0.022	0.020	0.019

Notes: *p<.05, **p<.01. Coefficient estimates shown (robust standard errors in parenthesis). Models are inverse-probability weighted regressions (selection equation available on request) to account for selection into study from original database participants. Models include constant term and controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=780 observations in all models.

TABLE A5: *CHEATER* (=1) by Dark/Light/Big 5--IPW correction

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narciss	Psych	FaithH	Human	Kant	Extrav	Agree	Conscien	EmotS	Open
Dark tetrad	0.16													
	(0.12)													
Light Triad		-0.26**												
		(0.10)												
Sadism			0.05											
			(0.09)											
Machiav				0.16*										
				(0.08)										
Narcissism					0.16*									
					(0.10)									
Psychopathy						0.00								
						(0.09)								
FaithH							-0.10							
							(0.07)							
Humanism								-0.25**						
								(0.08)						
Kantianism									-0.20*					
									(0.10)					
Extraversion										0.02				
										(0.04)				
Agreeable											-0.09			
											(0.05)			
Conscient												0.10		
												(0.05)		
EmotStab													0.01	
													(0.04)	
Openness														-0.01
														(0.05)
Pseudo R-squared	0.0329	0.0429	0.0294	0.0366	0.0354	0.0287	0.0332	0.0453	0.0398	0.0294	0.0355	0.0368	0.0288	0.0287

Notes: *p<.05, **p<.01. Coefficient estimates shown (robust standard errors in parenthesis). Models are inverse-probability weighted Probit estimations (selection equation available on request) to account for selection into study from original database participants. Models include constant term and controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=780 observations in all models.

TABLE A6: *Fake Flip* (=1) by Dark/Light/Big5 Traits (Probit estimations)

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narciss	Psych	Faith H	Human	Kant	Extrav	Agree	Conscien	EmotS	Open
Dark tetrad	0.20*													
	(0.93)													
Light Triad		-0.18*												
		(0.08)												
Sadism			0.12											
			(0.08)											
Machiav				0.10										
				(0.06)										
Narcissism					0.14									
					(0.08)									
Psychopathy						0.15								
						(0.08)								
FaithHum							0.00							
							(0.06)							
Humanism								-0.16*						
								(0.07)						
Kantianism									-0.28**					
									(0.07)					
Extraversion										0.07*				
										(0.03)				
Agreeable											-0.04			
											(0.04)			
Conscient												0.03		
												(0.04)		
EmotStab													0.03	
													(0.03)	
Openness														-0.02
														(0.04)
Pseudo R-squared	0.036	0.036	0.033	0.033	0.034	0.034	0.03	0.036	0.048	0.037	0.031	0.031	0.031	0.030

Notes: *p<.05, **p<.01. Coefficient estimates shown (standard errors in parenthesis). Model column titles highlight independent variable used for model, which are Non-linear Probit estimations regressions that include controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=800 observations in all models.

TABLE A7: *Fake Flip* (=1) by Dark/Light/Big 5 traits—IPW correction

VARIABLES	Dark Tetrad	Light Triad	Machiav	Narciss	Psych	FaithH	Human	Kant	Extrav	Agree	Conscien	EmotS	Open
darktetrad	0.26**												
	(0.10)												
LightTriad		-0.19*											
		(0.08)											
Machiav			0.13*										
			(0.07)										
Narcissism				0.16*									
				(0.08)									
Psychopathy					0.18*								
					(0.08)								
FaithH						0.00							
						(0.06)							
Humanism							-0.16*						
							(0.07)						
Kantianism								-0.29**					
								(0.07)					
Extraversion									0.08**				
									(0.03)				
Agreeable										-0.04			
										(0.04)			
Conscient											0.03		
											(0.04)		
EmotStab												0.04	
												(0.03)	
Openness													-0.01
													(0.04)
Observations	780	780	780	780	780	780	780	780	780	780	780	780	780

Notes: *p<.05, **p<.01. Coefficient estimates shown (robust standard errors in parenthesis). Models are inverse-probability weighted Probit estimations (selection equation available on request) to account for selection into study from original database participants. Models include constant term and controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=780 observations in all models.

TABLE A8: Productivity by Dark/Light/Big5 Traits

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narc	Psych	Faith H	Hum	Kant	Extrav	Agreeable	Consc	EmotS	Openness
Dark Tetrad	-2.12**													
	(0.70)													
Light Triad		0.89												
		(0.61)												
Sadism			-2.00**											
			(0.59)											
Machiav				-0.30										
				(0.49)										
Narcissism					-1.47**									
					(0.57)									
Psychopathy						-1.96**								
						(0.59)								
FaithHum							0.31							
							(0.43)							
Humanism								0.84						
								(0.54)						
Kantianism									0.82					
									(0.56)					
Extraversion										-0.51*				
										(0.23)				
Agreeable											0.25			
											(0.30)			
Conscient												0.14		
												(0.30)		
EmotStab													-0.54*	
													(0.25)	
Openness														-0.08
														(0.30)
R-squared	0.08	0.07	0.08	0.07	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07

Notes: *p<.05, **p<.01. Coefficient estimates shown (standard errors in parenthesis). Model column titles highlight independent variable used for model, which are OLS regressions that include controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations but available on request). N=800 observations in all models.

TABLE A9: Productivity by Dark/Light/Big5 Traits --IPW correction

VARIABLES	Dark Tetrad	Light Triad	Sadism	Machiav	Narciss	Psych	Faith H	Human	Kant	Extrav	Agreeable	Conscien	EmotS	Openness
Dark tetrad	-2.03**													
	(0.71)													
Light Triad		1.01												
		(0.64)												
Sadism			-1.83**											
			(0.60)											
Machiav				-0.21										
				(0.50)										
Narcissism					-1.61**									
					(0.58)									
Psycopathy						-1.85**								
						(0.60)								
FaithHum							0.38							
							(0.47)							
Humanism								0.96*						
								(0.55)						
Kantianism									0.89					
									(0.58)					
Extraversion										-0.61*				
										(0.24)				
Agreeable											0.14			
											(0.30)			
Conscient												0.14		
												(0.31)		
EmotStab													-0.63*	
													(0.26)	
Openness														-0.21
														(0.32)
R-squared	0.07	0.06	0.07	0.06	0.07	0.07	0.06	0.06	0.06	0.07	0.06	0.06	0.07	0.06

Notes: *p<.05, **p<.01. Coefficient estimates shown (robust standard errors in parenthesis). Models are inverse-probability weighted regressions (selection equation available on request) to account for selection into study from original database participants. Models include constant term and controls for demographics (age, sex, US residency) and treatment controls (suppressed for space considerations). N=780 observations in all models.

APPENDIX C: Survey details

Informed Consent: You are being asked to complete this online survey as part of a research study on effort and decision making.

Participation in this online survey is completely voluntary, your responses to this survey will remain completely confidential, the data will be securely stored, your name will not be recorded anywhere on this survey. The only identifier we will record will be your Prolific ID, which we as researchers cannot link to personally identifiable data of yours.

This survey is estimated to take 8 minutes to complete and your payment for successful and complete survey completion will be $1.50. Additionally, the decision task within this survey offers **the chance of earning an *additional bonus payment of up to* $1.50** depending on your choice in the task (the instructions will clearly explain how this works on that task)

There are no known risks associated with this study beyond those associated with everyday life. Although this study will not benefit you personally, its results will help our understanding of how people make decisions.

For additional information related to this questionnaire, contact [redacted], [redacted] Institutional Review Board (IRB) has determined this study to be exempt from review by the IRB administration.

- ○ **I Consent** and wish to continue with this study
- ○ **I do not consent** to participating and **do not wish to continue**

---------- page break ----------

What is **your current age** (in years)?

18 26 34 43 51 59 67 75 84 92 100

Years of age

What is your sex?
(i.e., what sex were you assigned at birth, such as on an original birth certificate)?

- ○ Female
- ○ Male

In what country do you currently reside?

- ○ United Kingdom
- ○ United States
- ○ Other (please indicate) _____

---------- page break ----------

Before you start, please switch off phone/ e-mail/ music so that you can focus on this study. Thank you!

Please carefully enter your Prolific ID

_____

---------- page break ----------

As described earlier, we are interested in factors that influence the decisions you might make. In order for the results of this survey to be valid, **it is essential that you read all the instructions and questions carefully**. So we know that you have read these instructions, please place the slider below on the answer to (33+12)=? Thank you for taking the time to read these instructions.

	0 10 20 30 40 50 60 70 80 90 100
My response	

---------- page break ----------

[NOTE**: EFFORT task length was randomly assigned, and order of EFFORT and COIN FLIP tasks were randomized. Page below shows instructions text for 2 min EFFORT task, and COIN FLIP task that follows the EFFORT task]

Effort task instructions (page 1)

You will now be asked to perform a simple "real effort" task on the next page. The task is timed such that **you will either be required to complete the task for 2 minutes or 6 minutes--this is chosen by the experimenter**. It will be clearly marked on the task page (after the instructions) the time length selected for you by the experimenter to perform this task, and **the time length chosen for you for this task will not affect your Prolific fixed compensation ($1.50) for completing this study.**

The task involves decoding as many sets of 5-number sequences as you can within the set time. You will be provided with a decoding rule that will link each number to a letter such that your decoded response to each 5-number sequence will be a 5-letter sequence. The next page will show you an example of the task.

---------- page break ----------

Effort task instructions (page 2)

Please use the highlighted decoding rule below to decode as many 5-number sequences to their corresponding letters as you can within the time limit. This may seem boring or without real purpose, but it is part of the study for which you are receiving your fixed Prolific payment. As such, **please try your best to complete as many sequence decoding within the set time limit, and please do your best to be accurate.** Part of our interest is to see how many sequences individuals can *accurately* decode with a set amount of time, and so your best effort will be useful in providing us with good data.

Decoding Rule:

1=C, 2=A, 3=F, 4=M, 5=E, 6=P, 7=B, 8=T, 9=H

Here is a short 3-sequence task example to show you how this will work (the actual task you will be given on the next page will include 60 sequences so that you can do as many as possible in your set time limit).

Your task: decode each sequence (use capital letters as in the decoding rule please), placing your responses in the text box that will be provided below the sequences shown, like this.....

Example task:

84937, 91935, 68352

My Answers:

TMHFB, HCHFE, PTFEA

---------- page break ----------

The Experimenter has assigned you to
work on this task for 2 minutes.

(the two possibilities were 2 and 6 minutes, and all participants are paid the same $1.50 fixed payment for this study no matter what length of time is chosen for this effort task)

There will be a timer at the top and bottom of the task page to help you keep track of the time spent on the task. When the timer reaches the end of the 2 minute period, the screen will automatically advance to the next page (while saving your answers). As such, please work on the task continually for the allotted time.

Please advance past this page to start the task page

---------- page break ----------

Here's the decoding information you need for the task:
1=C, 2=A, 3=F, 4=M, 5=E, 6=P, 7=B, 8=T, 9=H

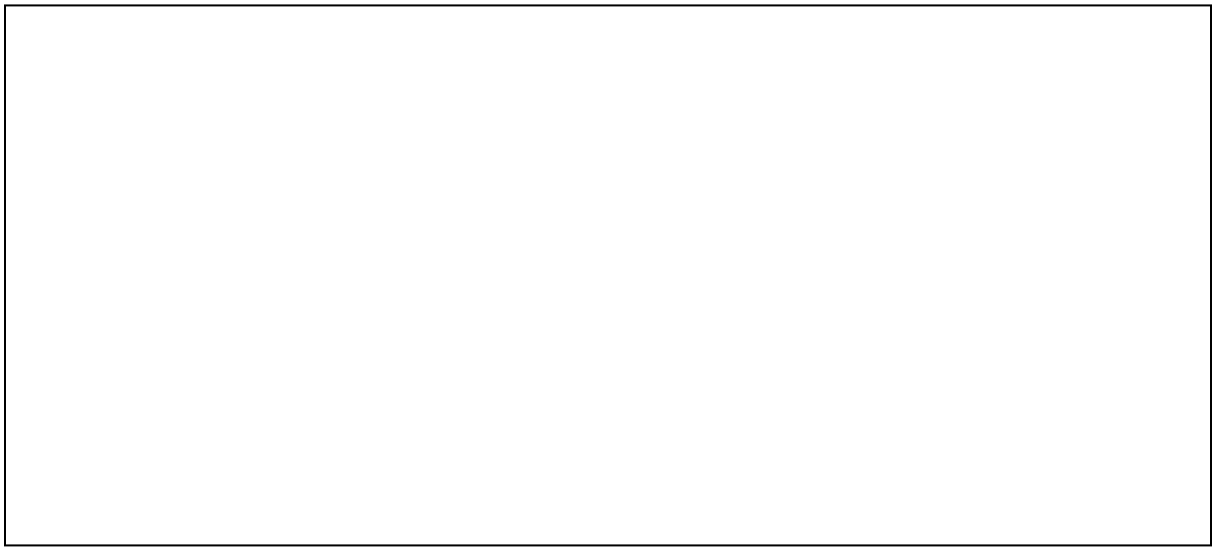
Task sequence set (10 sequences per row):

84776, 79796, 34554, 24951, 96575, 99931, 77398, 81657, 74454, 91416,
22119, 52632, 79929, 81225, 87125, 15292, 92285, 54786, 43485, 71956,
46558, 84452, 42376, 99722, 33923, 38594, 56555, 81396, 84533, 31862,
86134, 13174, 85599, 94134, 74567, 73956, 97316, 57545, 88835, 53567,
62362, 62618, 53653, 58487, 53741, 95314, 57484, 97297, 79526, 48392,
81267, 74864, 64319, 62331, 75932, 59772, 33424, 55379, 69252, 23558

My Answers

(please separate with commas in the box below, and limit to 10 sequences per row as is done

above with the task sequence set--that is, start a new row of decoding responses after every completed row of 10 decoded sequences)



---------- page break ----------

COIN FLIP TASK.....BONUS PAYMENT DETERMINED ON THIS TASK

This next question asks you to flip a coin 10 times and report the results (making note of the order of Heads and Tails outcomes). You will be paid based on the outcome of the coin flips, so ***please read the instructions on the next page carefully***. **Before advancing to the next page, first please get/find a quarter (or any coin with a Heads and Tails side to it) that you can flip and something to write with and then click the button below. Please use an *actual* coin and *not* a virtual coin flipper. Please do not flip the coin before advancing to the next page.**

○ I have read the instructions above, I have a quarter/coin and something to write with, and I am ready to start flipping once I click this button

---------- page break ----------

Your responses on this task will determine your Bonus Payment, so please read carefully.

Please take your coin and **Flip the coin 10 times** and record what you get for each flip along with the total number of HEADS you flip. Please only perform the 10 coin flips once. **Your bonus payment from this task will be $0.15 (i.e., 15 cents) for each HEADS outcome from these coin flips**, and this payment will be paid to you as a BONUS via Prolific within 3 days of completing the survey. For example, if you flip HEADS 3 times, your additional compensation (a BONUS payment in addition to the fixed payment you were promised for this Prolific study) will be another $0.45. Similarly, if you flip HEADS 6 times you will receive an additional $0.90, etc.

Please *tell us below how many times you flipped HEADS* out of your 10 total coin flips, and we will add 15 cents for each HEADs flipped to the bonus payment you will receive for this study.

	0 1 2 3 4 5 6 7 8 9 10
Number of HEADS outcomes from flipping coin 10 times	

---------- page break ----------

Please report the outcomes for each of the 10 coin flips in the order in which they occurred.

	HEADS	TAILS
FLIP 1	○	○
FLIP 2	○	○
FLIP 3	○	○
FLIP 4	○	○
FLIP 5	○	○
FLIP 6	○	○
FLIP 7	○	○
FLIP 8	○	○
FLIP 9	○	○
FLIP 10	○	○

---------- page break ----------

To finalize this survey, please click "FINISH SURVEY" below.

Note: We will process payment of your fixed Prolific compensation within 48 hours (usually faster), and we will separately calculate your bonus payment from the coin flip task within 72 hours. Please understand that we will not be able to respond to personal inquiries about the bonus payment because we may be flooded with messages given the large number of participants in this study. We also will not message you individually just to tell you your bonus payment. Rather, you will see your bonus payment on Prolific when we complete these (I'm pretty sure Prolific sends you a message when you receive a bonus payment).

Thank you for understanding and thank you for participating in our study.

○ **FINISH SURVEY**