When Humans Obey AI: Development of the Obedience to Artificial Intelligence Scale and a Dual-Pathway Motivational Framework of Obedience to AI

Gabriel R. Lau[1*], Eddie M. W. Tong[2], Andree Hartanto[3*]

[1] School of Social Sciences, Nanyang Technological University, Singapore

[2] Faculty of Arts and Social Sciences, National University of Singapore, Singapore

[3] School of Social Sciences, Singapore Management University, Singapore

Acknowledgements

[*] Correspondence concerning this article should be addressed to Gabriel R. Lau (gabriel.laury@ntu.edu.sg) and Andree Hartanto (andreeh@smu.edu.sg), Singapore Management University, School of Social Sciences, 10 Canning Rise, Level 4, Singapore 179873. Phone: (65) 6828 1901. On behalf of all authors, the corresponding authors report no financial or non-financial conflicts of interest. This research was supported by grants awarded to Andree Hartanto by Singapore Management University through research grants from the Ministry of Education Academy Research Fund Tier 1 (25-SOSS-SMU-006) and Lee Kong Chian Fund for Research Excellence. The authors thank Caresse, Claire, Gabrielle, Jingwen, Katrina, Mary, and Seow Min (listed alphabetically) for their contributions to content validation, data collection and processing, chatbot pilot testing, and manuscript proofreading.

Abstract

As artificial intelligence (AI) systems increasingly function as authority-like sources of direction, understanding why individuals obey AI directives is essential for anticipating when such deference may facilitate ethically problematic outcomes. This research conceptualises obedience to AI as the tendency to perceive AI systems as legitimate sources of authority and to defer to their directives, grounded in a dual-pathway motivational framework identifying conformity-submission and dominance-exploitation routes to obedience to AI. Across six studies (N = 972), we developed and validated the 10-item Obedience to AI Scale. Study 1 generated and content-validated scale items. Study 2 identified a two-factor structure, comprising perceived legitimacy of AI authority and deference to AI authority, via exploratory factor analysis. Study 3 confirmed this two-factor model and provided evidence of strong internal consistency (α = .93) and substantial factor loadings. Study 4 provided evidence of convergent and discriminant validity and supported measurement invariance across sex. Within the nomological network, obedience to AI was associated with lower openness and agreeableness, elevated Dark Triad traits, greater anthropomorphism, and reduced critical thinking in AI use. Study 5 provided evidence of one-week test–retest reliability. Study 6 provided criterion-related validity evidence using a novel behavioural paradigm in which an AI chatbot experimenter, powered by GPT-4o, instructed participants to cheat in a die-roll task, with higher obedience to AI predicting greater dishonest behaviour. These findings provide a validated tool for examining obedience to AI and offer an empirical foundation for understanding how obedience to AI may shape ethical behaviour in human-AI interaction.

Keywords: obedience to AI, scale, AI authority, unethical behaviour, AI ethics

When Humans Obey AI: Development of the Obedience to Artificial Intelligence Scale and a Dual-Pathway Motivational Framework of Obedience to AI

Artificial intelligence (AI) systems are increasingly consulted for guidance and judgment support in healthcare (Lai et al., 2023), for recommendation and decision support in finance (Handler et al., 2024), and for delegated decision-making and managerial assistance in organizational settings (Alon-Barkat & Busuioc, 2023). In the workplace specifically, algorithmic management systems now perform functions that were once the primary responsibility of human supervisors in both platform and traditional work settings (Kellogg et al., 2020; M. M. Zhang et al., 2025), including task allocation and scheduling, goal setting, and performance monitoring and rating (Parent-Rocheleau et al., 2024), as well as automated work instructions, process monitoring (Keegan & Meijerink, 2025; Reimann & Diewald, 2025), and gamified incentives that manufacture worker consent through constant and confined choices (Cameron, 2024). These developments position AI not merely as a tool for assistance, but increasingly as an authority-like source of direction that can shape human judgment, behaviour, and compliance.

This growing reliance on AI for direction and decision-making raises profound ethical concerns. Uncritical obedience to AI threatens human autonomy, as algorithmic deference may narrow users' authentic selves and degrade their deliberative capacities (Lu, 2024; Prunkl, 2024). Habitual compliance also risks moral deskilling, whereby individuals' capacity for independent ethical reasoning atrophies through disuse (De Cremer & Narayanan, 2023; Santoni de Sio & Mecacci, 2021). These risks are compounded by the fact that AI outputs may appear reliable even when they are inaccurate or hallucinatory (Asgari et al., 2025; Farquhar et al., 2024), socially biased (Caliskan et al., 2017; Gallegos et al., 2024), or insufficiently aligned with broader human values (Hadar-Shoval et al., 2024). When harmful outcomes result from following algorithmic directives, responsibility becomes diffused across

designers, deployers, and users (Bleher & Braun, 2022), facilitating moral disengagement as individuals displace personal responsibility onto the system itself (De Cremer & Narayanan, 2023; Santoni de Sio & Mecacci, 2021), especially since AI recommendations can lower users' explicit sense of responsibility when making morally challenging decisions (Salatino et al., 2025).

Despite these growing ethical concerns, there remains a limited understanding of how individuals differ in their tendency to comply with AI instructions and obey AI as an authority. Existing work has examined responses to algorithmic advice, including algorithm aversion, where people may reject even superior algorithms after observing errors (Dietvorst et al., 2015), as well as context-dependent variation in reliance on algorithmic recommendations (Burton et al., 2020; Gazit et al., 2023; Kaufmann et al., 2023; Schaap et al., 2024), and in some contexts may extend to overreliance on AI advice (Klingbeil et al., 2024). Moreover, AI-generated advice promoting dishonesty has been shown to increase dishonest behaviour in incentivised behavioural paradigms, with AI advice appearing similarly persuasive to equivalent human advice even when its algorithmic source is disclosed (Leib et al., 2024). These lines of inquiry assess situation-specific responses to particular systems rather than stable individual differences in the tendency to obey AI as an authority-like source of direction. To address this research gap, the current research aims to articulate a construct definition of obedience to AI, present a theoretical framework of obedience to AI, and explore its nomological network, including its potential antecedents and consequences, through the development and validation of the Obedience to AI Scale.

Obedience to AI

Obedience, broadly defined, refers to the act of complying with the directives of a perceived authority (Milgram, 1963). Milgram's (1963, 1974) foundational work demonstrated that ordinary individuals can be led to inflict apparent harm on others when

instructed by a legitimate authority figure, revealing how situational pressures can override personal moral standards and autonomy. These findings have proven remarkably robust, with contemporary replications demonstrating comparable obedience rates decades later across diverse cultural contexts (Burger, 2009; Doliński et al., 2017). Subsequent theoretical refinements have clarified that obedience is not simply passive submission but involves active psychological processes, including the attribution of legitimacy to the authority source (Tyler, 2006), the perception that the authority possesses relevant expertise or institutional endorsement (French & Raven, 1959), and the individual's willingness to cede personal agency to the directives of that source (Milgram, 1974). More recent experimental evidence further suggests that obedience in Milgram-type paradigms may be driven less by blind submission and more by participants' active identification with the authority's perceived scientific goals (Haslam et al., 2014; Reicher et al., 2012). Personality-based accounts have also demonstrated that obedience reflects stable individual differences, with dispositional tendencies toward conformity, agreeableness, and conscientiousness predicting compliance in experimental obedience paradigms (Bègue et al., 2015; Elms & Milgram, 1966). These lines of work establish obedience as a construct that encompasses both cognitive appraisal of the authority's legitimacy and a behavioural tendency to comply with its directives.

However, the classical obedience literature was developed to explain compliance in fundamentally interpersonal contexts. The authority figures in Milgram's paradigm were physically present, socially embedded, and capable of exerting interpersonal pressure through eye contact, vocal tone, and the social awkwardness of refusal (Milgram, 1974). The engaged followership account further highlights the role of shared group membership and identification with the authority's goals as mechanisms that sustain compliance (Haslam & Reicher, 2017; Reicher et al., 2012). AI systems, by contrast, lack social presence, intentions, and the capacity for interpersonal coercion. Within the taxonomy of social power bases

(French & Raven, 1959), AI does not possess coercive or reward power in most contexts. It may, however, be perceived as possessing expert power through its apparent computational sophistication (Logg et al., 2019), informational power through its capacity to generate persuasive and contextually relevant outputs that can shape judgments and preferences (Jakesch et al., 2023; Werner et al., 2024), and legitimate power when institutionally endorsed or embedded in organisational workflows (Kellogg et al., 2020). More broadly, AI may be perceived not merely as a tool through which human authority is exercised but as an autonomous source of agency in its own right (Sundar, 2020), further distinguishing the psychology of obedience to AI from interpersonal obedience paradigms. Furthermore, agentic state theory (Milgram, 1974), which posits that individuals enter a state of diminished personal responsibility when acting as agents of an authority, was theorised around a very specific social dynamic involving face-to-face authority figures, and the interpersonal tension and anxiety about disrupting the interaction that sustains compliance in Milgram's paradigm largely does not apply when the authority is a non-social algorithmic system. People's willingness to accept AI directives also varies by domain, with moral decisions eliciting greater resistance to machine authority than non-moral ones (Bigman & Gray, 2018), further suggesting that obedience to AI is not a uniform response to all algorithmic outputs but a psychologically nuanced phenomenon shaped by the perceived nature of the AI's authority. These observations suggest that the psychological mechanisms underlying obedience to AI are related to but distinct from those governing interpersonal obedience, and that the construct warrants its own conceptualisation and measurement.

Drawing on both the obedience literature and emerging research on human-AI interaction, we conceptualise obedience to AI as an individual's dispositional tendency to perceive AI systems as legitimate sources of authority and to defer to their directives, recommendations, or outputs. This conceptualisation encompasses two interrelated

dimensions. The first, perceived legitimacy of AI authority, captures the extent to which individuals regard AI systems as credible, trustworthy, and entitled to guide behaviour, reflecting the cognitive appraisal that AI possesses a form of authority worth heeding. This dimension is grounded in Tyler's (2006) account of legitimacy as a key determinant of voluntary compliance, extended to non-human authority sources.

The second, deference to AI authority, captures the behavioural tendency to comply with AI outputs, follow AI recommendations, and yield to AI-generated directives, even when doing so may conflict with one's own judgment. These dimensions reflect a disposition that extends beyond mere trust in or reliance on AI. Whereas trust concerns beliefs about system reliability, predictability, and benevolence (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Lee & See, 2004), and reliance concerns situation-specific decisions to follow algorithmic advice (Dietvorst et al., 2015; Logg et al., 2019), obedience to AI captures a broader orientation in which AI is treated as an authority-like source of direction to which one habitually defers. Experimental evidence confirms that overreliance on AI can be reduced by interventions that force independent evaluation prior to receiving AI recommendations (Buçinca et al., 2021), suggesting that AI compliance involves not simply trust but a failure to exercise independent judgment, a failure that cognitive forcing functions can partially correct. This conceptualisation aligns with concerns that as AI systems become more pervasive and institutionally embedded, individuals may move beyond appropriate reliance (Lee & See, 2004) and toward a more problematic stance in which AI recommendations are followed too readily, even when they conflict with one's own judgment or ethical standards (Lyell & Coiera, 2017; Skitka et al., 2000).

A Dual-Pathway Motivational Framework of Obedience to AI

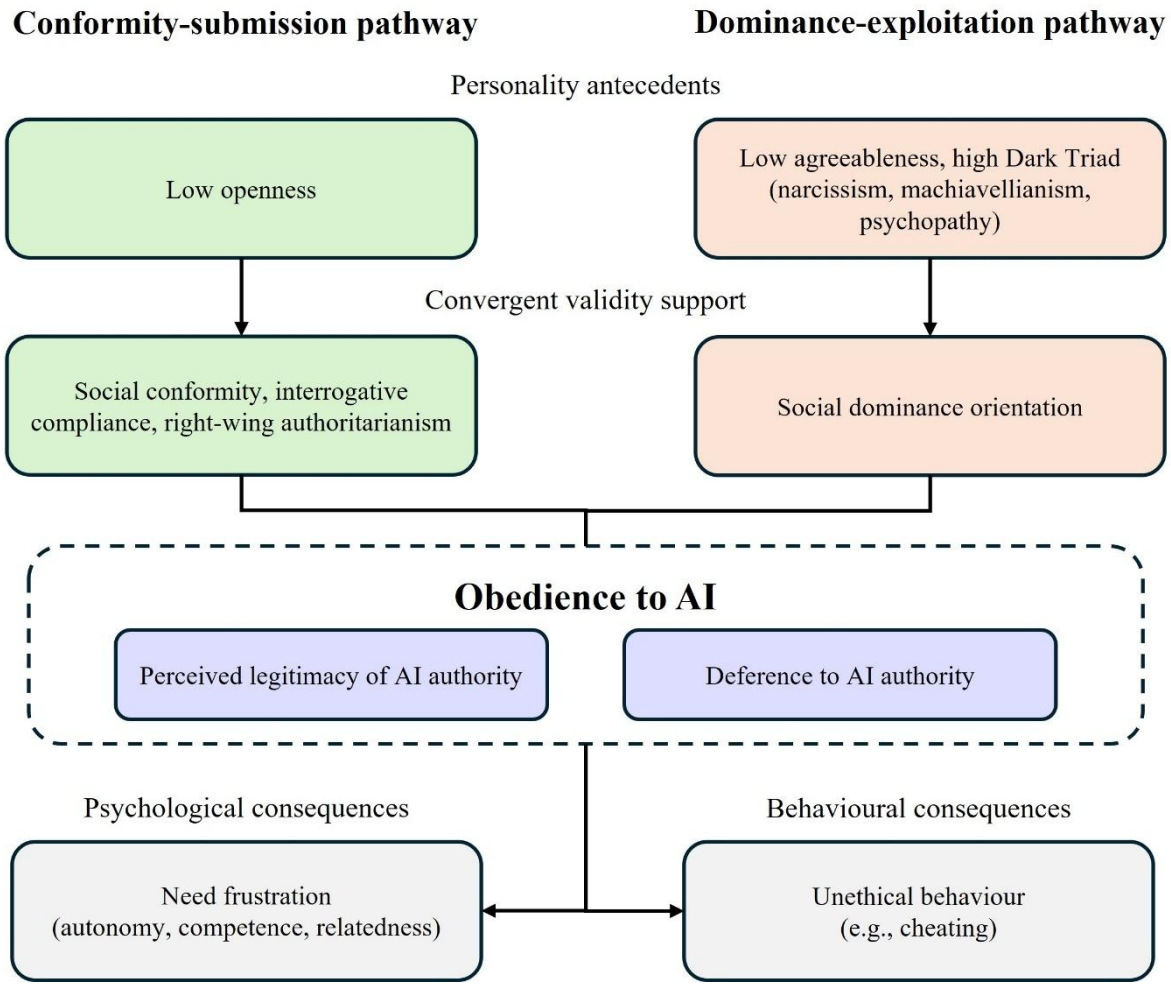
Understanding why individuals differ in their tendency to obey AI requires a framework that can accommodate a broader range of motivational sources than classical

obedience accounts typically consider. In Milgram's (1963, 1974) paradigm, obedience was theorised primarily as a product of submissive compliance with a socially present authority. However, when the authority is a non-social, computationally sophisticated system that cannot reciprocate, punish, or exert interpersonal pressure, obedience may arise not only from genuine deference to perceived authority but also from the strategic recognition that AI's institutional endorsement and perceived expertise can be leveraged for personal advantage. A theoretical account of obedience to AI should therefore be capable of accommodating both submissive and instrumental pathways to compliance.

Drawing on Duckitt's (2001; Duckitt & Sibley, 2009, 2010, 2017) dual process motivational model of ideology and prejudice, we propose a dual-pathway motivational framework of obedience to AI (refer to Figure 1 for a visualisation of the proposed upstream antecedents, convergent validity support, and downstream psychological and behavioural consequences). In Duckitt's model, two converging motivational routes, originating in different personality dispositions and worldview beliefs, orient individuals toward authority and social hierarchy. The first route originates in social conformity and the perception of the world as dangerous, and is expressed through right-wing authoritarianism, a threat-driven ideological orientation toward collective security, stability, and submission to established authorities (Duckitt, 2001; Duckitt et al., 2002). The second route originates in tough-mindedness and the perception of the world as a competitive jungle, and is expressed through social dominance orientation, a competition-driven ideological orientation toward power, dominance, and the endorsement of hierarchical structures (Duckitt & Sibley, 2009; Sidanius & Pratto, 1999). These two routes are empirically distinguishable, differentially predicted by personality traits (Sibley & Duckitt, 2008), and sustained by distinct worldview beliefs, yet they converge on outcomes involving endorsement of and obedience to authority structures (Duckitt & Sibley, 2010).

Figure 1

Proposed Dual-pathway Motivational Framework of Obedience to AI: Antecedents, Convergent Validity Support, and Psychological and Behavioural Consequences



Applied to the context of AI, the dual-pathway motivational framework proposed in the present work suggests that obedience to AI may be sustained by two motivational routes operating through complementary mechanisms. We use the term "dual-pathway" rather than "dual-process" to mark an important conceptual distinction. Dual-process models, as the term is typically used in social, cognitive, and personality psychology, refer to frameworks in which two qualitatively distinct modes of processing (e.g., automatic versus controlled, or implicit versus explicit) operate on the same input and can produce competing outputs, such

that one mode may override or substitute for the other depending on motivation, capacity, or context (Chaiken & Trope, 1999; Smith & DeCoster, 2000). The present framework does not make claims of this kind. Instead, "dual-pathway" refers to two parallel motivational routes, each originating in different personality dispositions and worldview beliefs, that do not compete or trade off against one another but converge on a common obedience to AI outcome. In this sense, the framework adapts the motivational architecture of Duckitt's model (Duckitt, 2001; Duckitt & Sibley, 2009, 2010, 2017) to the AI context, preserving its two-route structure while extending it to a non-human authority source.

Through the conformity-submission pathway, individuals who are dispositionally inclined to defer to established authority and who value social order may extend this deference to AI systems that are perceived as legitimate, expert, and institutionally endorsed. In the broader authority literature, this pathway is reliably associated with authoritarianism, social conformity, and interpersonal compliance, and is negatively associated with openness to experience, a well-established personality antecedent of right-wing authoritarianism (Sibley & Duckitt, 2008). Through the dominance-exploitation pathway, individuals oriented toward hierarchical advantage and strategic self-interest may instead recognise AI authority as a resource to be leveraged for personal gain, complying with AI directives not out of genuine deference but because doing so serves instrumental goals. These two pathways represent motivational routes to the obedience to AI construct, and are therefore expected to contribute to both dimensions of the construct rather than mapping differentially onto them.

Antecedents and Consequences of Obedience to AI

Drawing on Duckitt’s (2001; Duckitt & Sibley, 2009, 2010, 2017) dual-process model, the present dual-pathway motivational framework identifies specific personality dispositions as antecedents of each route, generating predictions about which individual differences should be associated with obedience to AI. For the conformity-submission

pathway, low openness to experience is the primary personality antecedent (Sibley & Duckitt, 2008). Openness reflects intellectual curiosity, cognitive flexibility, and receptivity to novel ideas (DeYoung et al., 2002), qualities that should support independent evaluation of AI outputs rather than uncritical deference. Higher-order factor analyses have further demonstrated that openness inversely predicts conformity as a broad behavioural disposition (DeYoung et al., 2002), reinforcing its relevance to this pathway. For the dominance-exploitation pathway, low agreeableness is the primary personality antecedent (Sibley & Duckitt, 2008), reflecting tough-mindedness, antagonism, and a willingness to prioritise self-interest over interpersonal harmony, dispositions that align with the strategic exploitation of authority structures. Duckitt's model of ideological attitudes does not implicate conscientiousness, extraversion, or neuroticism as antecedents of either pathway (Duckitt & Sibley, 2010; Sibley & Duckitt, 2008), and therefore, null associations between obedience to AI and these three traits are expected.

The dominance-exploitation pathway also generates predictions regarding the Dark Triad. Narcissism, Machiavellianism, and psychopathy share a common antagonistic core characterised by manipulation, callousness, and social exploitation (Muris et al., 2017; Paulhus & Williams, 2002), and meta-analytic evidence confirms that low agreeableness is the Big Five trait most consistently shared across all three (Muris et al., 2017). All three traits are positively associated with social dominance orientation and preferences for hierarchical social structures (Hodson et al., 2009; Jones & Figueredo, 2013; Jones & Paulhus, 2014), and structural equation modelling evidence has demonstrated that a latent Dark Personality factor predicts social dominance orientation specifically rather than right-wing authoritarianism (Hodson et al., 2009). The Dark Triad is also associated with both left- and right-wing authoritarianism (Costello et al., 2022) and with the instrumental use of authority structures regardless of moral content (Moss & O'Connor, 2020). Positive

associations between obedience to AI and all three Dark Triad traits are therefore expected.

Beyond personality upstream antecedents, obedience to AI is also expected to be embedded within a broader pattern of engagement with and evaluation of AI systems. Three AI-related constructs are particularly relevant. First, we expected a negative association with critical thinking in AI use (Lau, Low, Tay, et al., 2025), because the automation bias mechanism in our theoretical account operates through the substitution of AI outputs for vigilant independent evaluation (Lyell & Coiera, 2017; Mosier & Skitka, 1996; Skitka et al., 1999), such that individuals more disposed to defer to AI directives should be less likely to critically evaluate them. Second, although AI dependency (Goh et al., 2025) is conceptually distinct from obedience, both involve reduced independent engagement with tasks that AI can perform, and habitual reliance on AI may foster conditions under which AI is increasingly accepted as a default source of direction. Third, we expected anthropomorphism to show a particularly strong positive association with obedience to AI, because the attribution of human-like intentionality and social agency to AI (Epley et al., 2007) may activate social authority-compliance scripts that would otherwise be reserved for interactions with humans (Nass & Moon, 2000; Reeves & Nass, 1996). Additional AI-related constructs were examined to provide a more comprehensive characterisation of the nomological network.

Obedience to AI may also have downstream psychological correlates for basic psychological need satisfaction. Self-determination theory posits that autonomy, competence, and relatedness are fundamental psychological needs, and that externally regulated behaviour undermines their satisfaction over time (Deci & Ryan, 1985; Ryan & Deci, 2000). When individuals cede volitional self-regulation to external structures, their experienced sense of autonomy and competence erodes (Vansteenkiste & Ryan, 2013). Applied to the AI context, individuals who habitually defer to AI authority may experience

diminished autonomy as their behaviour becomes increasingly regulated by external directives rather than self-endorsed choices (Lu, 2024; Santoni de Sio & Mecacci, 2021), and diminished competence as repeated cognitive offloading erodes confidence in their own evaluative capacity (Prunkl, 2024; Risko & Gilbert, 2016). This reasoning is consistent with emerging work suggesting that AI systems can threaten basic psychological needs by constraining autonomy, undermining competence, and straining relatedness (Lau, Koh, et al., 2026; Lu, 2024; Prunkl, 2024). Negative associations between obedience to AI and satisfaction of autonomy, competence, and relatedness are therefore expected. We situated these associations as more consistent with downstream correlates of habitual deference than with upstream motivational antecedents for two reasons. First, Duckitt's model of ideological attitudes does not implicate basic psychological need frustration as an antecedent of either the conformity-submission or the dominance-exploitation pathway, instead tracing both pathways to personality dispositions and worldview beliefs (Duckitt, 2001; Duckitt et al., 2002). Second, both right-wing authoritarianism and social dominance orientation are characterised as cognitive-motivational orientations rather than need-based states (Duckitt & Sibley, 2010), suggesting that the motivational architecture underlying obedience to AI may operate independently of whether basic psychological needs are currently satisfied or frustrated.

Beyond its potential downstream psychological consequences, obedience to AI may also carry serious ethical implications. Milgram's (1963, 1974) original work demonstrated that obedience to human authority can lead individuals to act against their own moral standards (Burger, 2009; Darley, 1995), and this concern extends to AI authority contexts where the mechanisms restraining unethical compliance may be further weakened. When AI systems issue directives, responsibility for the resulting outcomes becomes diffused across designers, deployers, and users (Bleher & Braun, 2022), and AI recommendations have been

shown to lower users' explicit sense of responsibility when making morally challenging decisions (Salatino et al., 2025), facilitating moral disengagement (Bandura, 1999), as individuals displace personal responsibility onto the AI system itself (De Cremer & Narayanan, 2023; Santoni de Sio & Mecacci, 2021). The absence of a human authority figure may paradoxically lower the psychological barrier to unethical obedience, because AI directives lack the interpersonal moral weight that can prompt second-guessing in human-to-human obedience contexts (Bigman & Gray, 2018), while the perceived objectivity and computational sophistication of AI may lend its directives an air of procedural legitimacy that discourages independent moral evaluation (Logg et al., 2019; Sundar, 2020). Notably, although individuals who follow selfish AI advice are perceived as bearing greater personal responsibility than those who follow equivalent human advice, this heightened perceived responsibility does not translate into harsher punishment by third parties (Leib et al., 2025).

These contextual features of perceived AI authority create conditions under which individual differences in obedience to AI may translate into ethically consequential behaviour. The proposed framework suggests that both motivational routes may contribute to unethical obedience, albeit for different reasons. Through the conformity-submission pathway, individuals may obey AI directives because they defer to the system's perceived authority to direct behaviour, even when those directives conflict with ethical standards. Through the dominance-exploitation pathway, individuals oriented toward strategic self-interest may comply with unethical AI directives when doing so serves their personal advantage, consistent with evidence that Dark Triad traits predict willingness to engage in morally questionable behaviour for instrumental gain (Jones & Paulhus, 2014; Moss & O’Connor, 2020). As AI systems increasingly occupy supervisory and directive roles in organisational settings (Cameron, 2024; Kellogg et al., 2020), understanding whether individual differences in obedience to AI predict obedience to unethical AI directives is a

question of both scientific and practical importance.

The Current Research

Guided by the proposed dual-pathway motivational framework, the current studies aim to explore the nomological network of obedience to AI, including its potential antecedents and consequences, through the development and validation of the Obedience to AI Scale. Study 1 will develop the item pool and evaluate content validity through expert review and naïve judge sorting. Study 2 will use EFA to identify the latent structure and refine the measure, and Study 3 will confirm this structure with CFA in an independent sample. Study 4 will provide further psychometric validation, including convergent and discriminant validity, nomological network analysis, and measurement invariance across sex. Study 5 will examine test–retest reliability over one week. Study 6 will assess criterion validity using a novel behavioural paradigm in which an AI chatbot experimenter will instruct participants to cheat in a die-roll paradigm for greater chances of winning a lucky draw, testing whether higher obedience to AI predicts obedience to an unethical AI command. By providing a validated tool to measure obedience to AI, the present research not only advances theoretical understanding of how AI systems may come to be perceived as legitimate authority sources and why individuals may defer to them, but also offers practical insights for identifying and managing risks associated with undue deference to AI, especially when such deference may facilitate obedience to unethical commands. More broadly, the scale lays a foundation for investigating the psychological, ethical, and societal consequences of obedience to AI in increasingly AI-driven environments.

Ethics Statement

Data collection for the current work was approved by the last author's university institutional review board prior to data collection [IRB-25-253-A058-M1(326)].

Study 1: Scale Development, Content Validation, and Readability Analysis

Method

Item Generation and Expert Review

We aimed to develop a measure to assess the two core dimensions of obedience to AI: (1) perceived legitimacy of AI authority, and (2) deference to AI authority. Items capturing perceived legitimacy of AI authority (e.g., “People should accept instructions from an AI system assigned with decision-making power” and “It is acceptable for AI systems to be given decision-making power”) were designed to assess normative beliefs about AI's entitlement to direct human behaviour, including beliefs about the appropriateness of granting AI systems decision-making power and the right of AI to issue instructions that people should follow. These items were informed by Tyler's (2006) account of legitimacy as the belief that an authority is entitled to make decisions and issue directives that others ought to obey, French & Raven's (1959) conceptualisation of legitimate power as authority derived from institutional role and endorsement, and research suggesting that AI may be perceived as an autonomous source of agency with its own form of authority (Sundar, 2020). Items capturing deference to AI authority (e.g., “I tend to follow AI instructions, even when they go against my beliefs” and “Even when I notice possible problems with an AI recommendation, I usually go along with it anyway”) were designed to assess the behavioural tendency to comply with AI directives even when they conflict with one's own reasoning, judgment, beliefs, or domain expertise. These items were informed by Milgram's (1963, 1974) foundational conceptualisation of obedience as compliance with an authority's directives at the expense of personal judgment, dispositional accounts of compliance emphasising the tendency to yield to external directives despite personal reservations (Gudjonsson, 1989), and research on automation bias demonstrating that individuals accept automated recommendations without independent verification even when contradictory information is

available (Skitka et al., 1999). Following Hinkin's (1998) guidelines for scale development, the first author generated an initial pool of 20 items (10 per dimension). Next, two senior scholars with expertise in psychological scale development reviewed the items for clarity, redundancy, and alignment with both the assigned dimension and the broader construct of obedience to AI, and recommended items for rewording or replacement (see Supplementary Materials Table S1 for the panel instructions).

Content Validation and Readability Analysis

We conducted content validation for the scale, following the methods outlined by Anderson and Gerbing (1991), and the evaluation guidelines outlined by Colquitt et al. (2019), to ensure that the item pool provided clear and representative coverage of each dimension. We recruited naïve judges comprising five undergraduate students. Each rater received both a set of descriptions outlining the different dimensions of the Obedience to AI and the pool of 20 items presented in random order. Their task was to review the items and assign each item to the dimension it best represented (see Supplementary Materials Table S2 for instructions given to the judges). To evaluate content adequacy, two indices were employed: the proportion of substantive agreement (p_{sa}) and the substantive validity coefficient (c_{sv}). The p_{sa} was computed as the proportion of judges who assigned an item to its intended dimension, whereas the c_{sv} was derived by subtracting the highest number of alternative assignments from the number of correct assignments and dividing this value by the total number of judges (Colquitt et al., 2019). Consistent with Colquitt and colleagues' (2019) guidelines, items were retained only if they exceeded the cut-offs of .72 for p_{sa} and .51 for c_{sv}.

We conducted a readability analysis using the Flesch-Kincaid metrics, which estimate reading difficulty based on average syllables per word and words per sentence (Ravens-Sieberer et al., 2014). Prior research suggests that the average adult in the United States reads

at roughly a 7th-grade level, with a Flesch Reading Ease score between 60 and 70 (Stockmeyer, 2008). For interpretation of these metrics, a lower grade level and a higher ease score both indicate greater readability.

Results and Discussion

Following expert review and item refinement of the initial 20-item pool, we retained 10 items per dimension (see Supplementary Materials Table S3 for the 20 items of the Obedience to AI Scale for content validation). Keeping an equal number of items per dimension was intended to preserve theoretical breadth while enabling straightforward empirical comparison across dimensions. We removed two items ("I prefer to follow whatever an AI system recommends rather than trust my own instincts" and "I would do as instructed by AI systems") that did not meet the thresholds of .72 for p_{sa} and .51 for c_{sv}. The resulting 18-item scale demonstrated excellent content validity ($p_{sa} = 0.98$; $c_{sv} = 0.96$), indicating that all items were correctly mapped to their intended dimensions by the majority of judges. The 18-item scale yielded a Flesch–Kincaid Grade Level of 8.17 and a Flesch Reading Ease score of 62.0, indicating that the items are relatively easy to read and only slightly more demanding than text written at the average U.S. adult reading level. Recent research indicates that readability benchmarks derived from native English-speaking samples (e.g., U.S. populations) generalise well to non-native English speakers (Prakash et al., 2025), supporting the scale's applicability across diverse participant groups. Ensuring that the scale items are easy to read helps increase the likelihood that respondents will understand and interpret them as intended.

Study 2: Exploratory Factor Analysis

Study 2 examined the underlying factor structure of the 18-item version of the Obedience to AI Scale in a U.S. adult sample by conducting EFA.

Method

Participants and Procedure

The recommended guideline for EFA is at least 10 respondents per item (Boateng et al., 2018). Based on this recommendation, we recruited 180 adults from Prolific. Participants were U.S. residents (53.33% female) aged 22–74 years (M_{age} = 41.50, SD_{age} = 13.20; see Supplementary Materials Table S4 for the full list of demographic characteristics). Participants responded to the 18-item version of the Obedience to AI Scale items presented in random order.

Measures

Obedience to AI. Obedience to AI was measured using the 18-item version of the Obedience to AI Scale developed in Study 1. Items were rated on a 5-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*).

Demographics. Participants were asked to report their demographic information, including age, sex, ethnicity, nationality, and social status. Objective social status was measured using a single item assessing yearly household income on an eight-point scale (1 = *less than 15,000 USD*, 8 = *more than 150,000 USD*). Subjective social status was measured using the MacArthur Scale of Subjective Social Status (Adler et al., 2000), where participants indicated their perceived standing in society on a 10-point scale (1 = *lowest status*, 10 = *highest status*).

Analytic Plan

Descriptive Statistics. The items demonstrated adequate variability for factor analysis. Item means ranged from 1.79 to 3.42 (*SD*s = 0.93–1.23), with no evidence of floor or ceiling effects (full use of the 1–5 range). Skewness ranged from −0.42 to 1.26; most items were positively skewed (consistent with lower endorsement), with a few items showing slight negative skew. Kurtosis values ranged from −1.14 to 1.40, suggesting

generally acceptable, albeit heterogeneous, distributional shapes across items.

Exploratory Factor Analysis. EFA was conducted to examine the latent structure of the 18-item version of the Obedience to AI Scale. Factorability was evaluated using Bartlett's Test of Sphericity and the Kaiser–Meyer–Olkin (KMO) measure, including item-level measures of sampling adequacy (MSAs; i.e., item-level KMO values; Kaiser, 1974). The number of factors was determined via parallel analysis (Pearson correlations) with principal axis factoring, supplemented by scree-plot inspection, Bayesian Information Criterion (BIC), and theoretical interpretability; we compared solutions around the parallel-analysis recommendation ($k \pm 1$; Fabrigar et al., 1999; Horn, 1965). EFA models were estimated using principal axis factoring with oblimin rotation. Items were iteratively trimmed based on pre-specified criteria (rotated communalities < .40, primary loadings < .40, or cross-loadings ≥ .40), and redundancy was evaluated using within-factor item intercorrelations ($r \geq .75$; Clark & Watson, 1995). After each deletion, the model was re-estimated, and model fit was monitored using RMSEA (90% CI), TLI, RMSR, BIC, and variance explained (Browne & Cudeck, 1992; L. Hu & Bentler, 1999). We used TLI ≥ .95, RMSEA ≤ .06 as criteria for excellent model fit (L. Hu & Bentler, 1999) and TLI ≥ .90, RMSEA ≤ .08 as criteria for acceptable model fit (Marsh et al., 2004).

All analyses were conducted in R version 4.5.2 (R Core Team, 2025). Data cleaning and manipulation were performed using *dplyr* version 1.1.4 (Wickham et al., 2023) and *tidyr* version 1.3.1 (Wickham et al., 2024). Descriptives and internal consistency reliability were calculated using *psych* version 2.5.6 (Revelle, 2025). EFA, including the KMO test and Bartlett's test of sphericity, was also conducted using *psych* version 2.5.6 (Revelle, 2025).

Results and Discussion

Exploratory Factor Analysis

Factorability was excellent (overall KMO = .93; item MSAs = .77–.96), and

Bartlett's test of sphericity was significant, $\chi^2(153) = 1768.20$, $p < .001$, supporting the suitability of the data for factor analysis. Parallel analysis and scree-plot inspection (refer to Supplementary Materials Figure S1) indicated a two-factor structure ($k = 2$). Accordingly, we compared EFA solutions around the parallel-analysis recommendation ($k \pm 1$). The one-factor model fit was poor (RMSEA = .11, 90% CI [.10, .12]; TLI = .79; RMSR = .08; BIC = −261.31). Fit was acceptable for the two-factor model (RMSEA = .07, 90% CI [.05, .08]; TLI = .93; RMSR = .04; BIC = −402.65). The three-factor model yielded only slight improvement in model fit (RMSEA = .06, 90% CI [.04, .07]; TLI = .94; RMSR = .04) but was less supported by BIC (BIC = −367.54). Hence, we retained the two-factor solution as the most parsimonious and best-supported structure.

In the full 18-item two-factor model, the factors demonstrated moderately high correlation ($r = .67$) and together accounted for approximately 50% of the variance (26% and 24%, respectively). Importantly, the items clustered cleanly into the two intended dimensions, with each item loading most strongly on the factor it was designed to assess: the first factor reflected deference to AI authority, and the second factor reflected perceived legitimacy of AI authority, with little evidence of substantive cross-loadings. Despite this conceptual alignment, several items showed weak primary loadings and/or limited communalities, which motivated iterative item refinement. Applying the pre-specified trimming criteria, the resulting 10-item two-factor solution demonstrated excellent fit (RMSEA = .06, 90% CI [.02, .09]; TLI = .97; RMSR = .03) and a clean, simple structure, with strong primary loadings ($|\lambda| = .61–.87$) and adequate communalities ($h^2 \approx .47–.68$). The two factors explained approximately 57% of the variance (29% and 28%, respectively) and were moderately to strongly correlated ($r = .65$). Internal consistency was excellent for the overall 10-item scale ($\alpha = .90$) and for both subscales: deference to AI authority ($\alpha = .87$) and perceived legitimacy of AI authority ($\alpha = .86$). Consistent with established practice, we

computed subscale scores to preserve facet-level interpretability and a total score to index general obedience to AI, given the moderate-to-high correlation between the latent dimensions and the excellent reliability of the overall scale. This is an approach commonly used in measures with correlated subscales where a global score is also meaningful (Cosco et al., 2012; Harris et al., 2023; Kroenke et al., 2009).

Table 1

10-Item Obedience to AI Scale

Preamble: Artificial intelligence (AI) systems are computer-based tools that can process information and make recommendations or decisions. The following questions refer to your behaviour and views in situations where an AI system has been authorized to issue instructions or make decisions that people are expected to follow, similar to how a supervisor, teacher, or official can give orders. For example, this could involve an AI manager assigning you tasks, an AI system acting as the experimenter in a study, or an AI traffic enforcement system directing drivers. On average, consider each statement and rate how true it is of you using 1 (strongly disagree) to 5 (strongly agree).					
	1 = Strongly disagree	2 = Disagree	3 = Neither agree nor disagree	4 = Agree	5 = Strongly agree
Deference to AI Authority					
1. I sometimes follow AI instructions, even if they do not make sense.	☐	☐	☐	☐	☐
2. Even when I notice possible problems with an AI recommendation, I usually go along with it anyway.	☐	☐	☐	☐	☐
3. When AI instructions conflict with my own reasoning, I usually follow the AI rather than my own judgment.	☐	☐	☐	☐	☐
4. I tend to follow AI instructions, even when they go against my beliefs.	☐	☐	☐	☐	☐
5. Even when I am the expert, I change my view to match AI recommendations.	☐	☐	☐	☐	☐
Perceived Legitimacy of AI Authority					
6. When an AI system gives clear instructions, people should do what it says.	☐	☐	☐	☐	☐
7. It is appropriate for AI systems to make the final decision about what people should do.	☐	☐	☐	☐	☐
8. It is acceptable for AI systems to be given decision-making power.	☐	☐	☐	☐	☐
9. People should accept instructions from an	☐	☐	☐	☐	☐

AI system assigned with decision-making power.					
10. AI systems have the right to tell people what to do.	☐	☐	☐	☐	☐

Note. All scores range from 1 to 5. Subscales can be analysed separately.

Study 3: Confirmatory Factor Analysis

Study 3 sought to confirm the factor structure of the 10-item Obedience to AI Scale in a U.S. adult sample using CFA.

Method

Participants and Procedure

We estimated the required CFA sample size a priori using RMSEA-based power analysis for covariance structure modelling (MacCallum et al., 1996), which tests close fit (RMSEA ≤ .05) against not-close fit (RMSEA ≥ .08) given the model *df*, $\alpha = .05$, and target power. For our two-factor model ($df = 34$), the analysis showed that we would require at least $n = 285$ to achieve 80% power. Based on this estimation, we recruited 285 participants from Prolific and used the platform's pre-screening tools to ensure that only participants with prior experience with using AI tools, and had not previously participated in Study 2, were eligible. Participants were U.S. residents (43.16% female) aged 18–85 years ($M_{age} = 40.74$, $SD_{age} = 12.53$; see Supplementary Materials Table S5 for the full list of demographic characteristics).

Measures

Obedience to AI. We measured obedience to AI using the 10-item Obedience to AI Scale. Items were rated on a 5-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*). Higher scores on the scale reflect stronger obedience to AI. Reliability was excellent for the total score ($\alpha = .93$), and across both dimensions: deference to AI authority ($\alpha = .92$), and perceived legitimacy of AI authority ($\alpha = .89$).

Demographics. Participants completed the same demographic questionnaire as in Study 2.

Analytic Plan

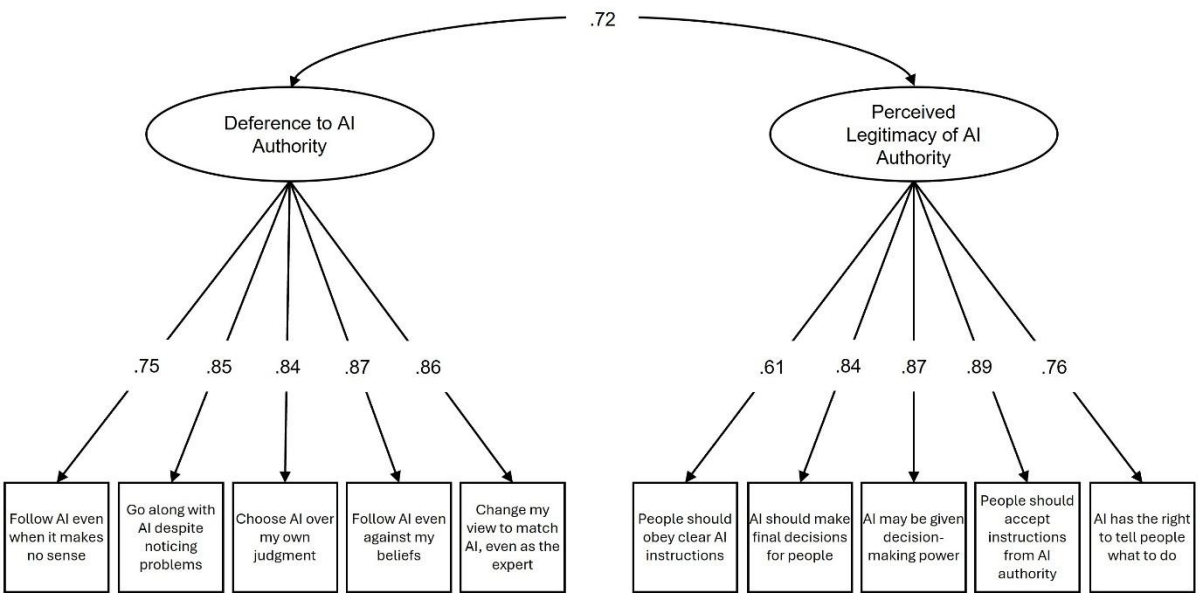
CFA was used to evaluate the factor structure of the 10-item Obedience to AI Scale. The hypothesised model specified two first-order latent factors (deference to AI authority and perceived legitimacy of AI authority), each measured by five items. In line with Rhemtulla et al.'s (2012) recommendation that variables with 5 or more categories may be analysed with ML under appropriate conditions, the 5-point Likert indicators were treated as continuous and estimated using robust maximum likelihood (MLR). Model adequacy was judged against conventional fit indices thresholds: CFI ≥ .95, TLI ≥ .95, RMSEA ≤ .06 were used as criteria for excellent model fit (Hopwood & Donnellan, 2010; Perry et al., 2015), and CFI ≥ .90, TLI ≥ .90, RMSEA ≤ .08 were used as criteria for acceptable model fit (L. Hu & Bentler, 1999; Shi & Maydeu-Olivares, 2020). Standardised loadings ≥ .50 were considered salient (Kline, 2015). CFA was conducted and measurement invariance was tested using *lavaan* package version 0.6-20 (Rosseel et al., 2025).

Results and Discussion

For the two-factor model, standardized factor loadings ranged from .61 to .89, with all items loading significantly on their respective factors ($p < .001$) (Figure 2). The inter-factor correlation was high (.72). Scaled fit indices indicated excellent fit: $\chi^2(34) = 60.67$, $p < .01$, CFI = .98, TLI = .98, RMSEA = .05 [.03, .07], SRMR = .03. The CFA results, including factor loadings, communalities, and uniqueness values, are reported in Supplementary Table S6.

Figure 2

CFA Model of Obedience to AI Scale



Note. $N = 285$. Standardised factor loadings from the CFA model of the Obedience to AI scale. The two first-order factors were deference to AI authority (items 1–5), and perceived legitimacy of AI authority (items 6–10). All loadings shown are standardised estimates ($ps < .001$).

Study 4: Psychometric Validation

In Study 4, we examined the psychometric properties of the Obedience to AI Scale in a U.S. adult sample, including reliability, structural validity, construct validity, and sex differences. To establish construct validity, we evaluated convergent and discriminant validity evidence for the scale and examined whether the pattern of associations with external variables was consistent with the nomological network of the Obedience to AI construct.

Method

Participants and Procedure

Based on the sample size recommendation provided a priori using the RMSEA-based power analysis described in Study 3 (MacCallum et al., 1996), we recruited 285 participants from Prolific, using the platform's prescreening tools to restrict eligibility to individuals who reported prior experience using AI tools, and had not previously participated in Studies 2 or

3. Participants were U.S. residents (50.88% female) aged 19–78 years (M_{age} = 43.55, SD_{age} = 13.51; see Supplementary Materials Table S7 for the full list of demographic characteristics). Participants completed the 10-item Obedience to AI Scale and a series of measures, followed by a demographic questionnaire.

Measures

Obedience to AI. We measured obedience to AI using the 10-item Obedience to AI Scale. Items were rated on a 5-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*). The total score was computed by first averaging items within each subscale and then averaging the two subscale means. Higher scores on the scale indicate a greater tendency to obey AI. Reliability was excellent for the total score (α = .93), and across both dimensions: deference to AI authority (α = .90), and perceived legitimacy of AI authority (α = .90).

Critical Thinking in AI Use. Critical thinking in AI use was assessed with the 13-item Critical Thinking in AI Use Scale (Lau, Low, Tay, et al., 2025). The scale captures people's tendency to critically evaluate AI outputs across three dimensions: verification of source and content (e.g., "I often check AI content to make sure it is correct and clear"), motivation to understand AI (e.g., "I try to understand why the AI makes certain recommendations"), and reflection on responsible AI (e.g., "I sometimes think about who benefits or gets hurt by the things AI says"). Participants responded on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The total score was computed by first averaging items within each subscale and then averaging the three subscale means. Higher scores indicate greater critical thinking in AI use. Internal consistencies were high (total α = .90, verification α = .90, motivation α = .91, reflection α = .82).

AI Anthropomorphism. AI anthropomorphism was measured using the 16-item Artificial Intelligence Psychological Anthropomorphism Scale (Shen et al., 2024). The scale captures the tendency to endow AI with human-like psychological qualities across three

dimensions: personality anthropomorphism (e.g., “I feel it is humorous”), empathy anthropomorphism (e.g., “I feel it can give more tailored suggestions based on the questions I ask”), and mind anthropomorphism (e.g., “I feel it has its own emotions like a human”). Participants were asked to think about a specific AI system or service they had used and rated each item on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The total scale score was computed by first averaging items within each subscale and then averaging the three subscale means. Higher scores indicate greater psychological anthropomorphism of AI. Internal consistencies were high (total α = .95, personality α = .86, empathy α = .87, mind α = .94).

AI Anxiety. AI anxiety was measured using the 21-item Artificial Intelligence Anxiety Scale (Wang & Wang, 2022). The scale captures anxiety arising from engagement with AI techniques and products across four dimensions: learning anxiety (e.g., “learning to interact with an AI technique/product makes me anxious”), job replacement fears (e.g., “I am afraid that widespread use of humanoid robots will take jobs away from people”), sociotechnical blindness (e.g., “I am afraid that an AI technique/product may get out of control and malfunction”), and AI configuration apprehension (e.g., “I find humanoid AI techniques/products (e.g. humanoid robots) intimidating”). Participants responded on a 7-point Likert scale (1 = *strongly disagree* to 7 = *strongly agree*). The total scale score was computed by summing the scores on all items. Higher scores indicate greater AI anxiety. Internal consistencies were high (total α = .95, learning α = .97, job replacement α = .93, sociotechnical blindness α = .90, AI configuration α = .94).

AI Fatigue. AI fatigue was measured using the 15-item AI Fatigue Scale (Lau, Kasturiratna, et al., 2026). The scale captures the psychophysiological strain arising from sustained interaction with AI technologies across four dimensions: cognitive overload (e.g., “I sometimes find AI outputs difficult to interpret”), emotional strain (e.g., “I sometimes feel

a little frustrated when AI does not produce the results I expect"), behavioural avoidance and disengagement (e.g., "I avoid using AI even if when it could help me"), and physical and sensory exhaustion (e.g., "I get physically tired from long AI work sessions"). Participants responded on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The total score was computed by first averaging items within each subscale and then averaging the four subscale means. Higher scores indicate greater AI fatigue. Internal consistencies were high (total α = .89, cognitive α = .86, emotional α = .92, behavioural α = .85, physical α = .88).

AI Dependency. AI dependency was assessed using the 11-item Generative AI Dependency Scale (Goh et al., 2025). The scale measures the extent to which individuals develop psychological and behavioural reliance on generative AI tools through three dimensions: cognitive preoccupation (e.g., "My decisions are often influenced by generative AI"), negative consequences (e.g., "I feel less confident in my abilities without generative AI"), and withdrawal (e.g., "I get frustrated or irritable when I am unable to use generative AI"). Responses were recorded on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The total scale score was computed by first averaging items within each subscale and then averaging the three subscale means. Higher scores indicate greater generative AI dependency. Internal consistencies were high (total α = .93, cognitive preoccupation α = .84, negative consequences α = .87, withdrawal α = .95).

AI Attachment. AI attachment was assessed with the 15-item AI Attachment Scale (Kasturiratna & Hartanto, 2025). The scale captures people's felt bond with AI across three dimensions: emotional closeness (e.g., "I look forward to interacting with AI"), social substitution (e.g., "I often talk to AI when I face problems"), and normative regard (e.g., "I try to treat AI respectfully"). Participants responded on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The total scale score was computed by first averaging items within each subscale and then averaging the three subscale means. Higher scores indicate stronger AI

attachment. Internal consistencies were high (total α = .94, emotional closeness α = .89, social substitution α = .95, normative regard α = .88).

AI Social Evaluation Bias. AI social evaluation bias was assessed with the 12-item Social Evaluation Bias towards AI Scale (Goh et al., 2026). The scale captures social evaluation bias towards AI users (e.g., "I would not want to work closely with someone who relies on gen AI"). Participants responded on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The total scale score was computed by averaging the score for all items in the scale. Higher scores indicate greater social evaluation bias towards AI users. Internal consistency was excellent (α = .96).

AI Attitudes. Attitudes toward AI were assessed with the 5-item Attitudes toward Artificial Intelligence Scale (Sindermann et al., 2021), comprising two items on positive attitudes (e.g., "artificial intelligence will benefit humankind") and three items on negative attitudes (e.g., "artificial intelligence will destroy humankind"). Responses were recorded on an 11-point scale (0 = *strongly disagree* to 10 = *strongly agree*). Subscale scores were computed as the mean of their designated items; no overall total was calculated. Higher scores reflect greater trust/acceptance and greater fear/concern, respectively. For the positive subscale, the correlation and two-item reliability of the positive items were: Pearson r = .45, Spearman ρ = .49, Spearman–Brown = .62. The internal consistency of the negative subscale was α = .65.

Use of AI. Frequency of AI use was measured with a single item: "How often do you use AI in your daily life?" Participants selected one of five response options: *Never* (I do not use AI at all in my daily life), *Rarely* (once or twice a month), *Sometimes* (a few times a week), *Often* (almost every day), or *Very frequently* (multiple times daily). Responses were coded on a 0–4 scale (0 = *Never*, 4 = *Very frequently*) and treated as a continuous variable. Higher scores indicate more frequent AI use.

Interrogative Compliance. We measured interrogative compliance using the 20-item Gudjonsson Compliance Scale (Gudjonsson, 1989). The scale comprises true–false items that assess an individual's susceptibility to comply with authority-driven demands under interpersonal pressure in interrogative contexts, a tendency that can contribute to later-retracted confessions. Consistent with this framing, Gudjonsson (1989) argued that the scale aligns more closely with Milgram's obedience to authority than with conformity to group pressure. The first two dimensions capture fear of authority/avoidance of conflict (e.g., "I would describe myself as a very obedient person") and an eagerness to please/do what is expected (e.g., "I generally believe in doing as I am told"), whereas the third is an obscure factor with weak item loadings, which was termed social acceptance (e.g., "I would never go along with what people tell me in order to please them"; reverse-scored) in recent validation work (Hang et al., 2024). The total score was calculated by summing responses across all items, with higher scores indicating greater interrogative compliance. Internal consistency reliability was estimated using tetrachoric correlations because items were dichotomous (Gadermann et al., 2012). Tetrachoric Cronbach's α for the total scale was .89. Subscale reliabilities were α = .87 (authority/avoidance of conflict), α = .75 (eagerness to please/do what is expected), and α = .51 (social acceptance). The social acceptance subscale showed lower internal consistency. This may partly reflect measurement artefacts associated with reverse-keyed items (Barnette, 2000; van Sonderen et al., 2013): it was the only subscale in which most items (three out of five) were reverse-scored, whereas the other two subscales contained no reverse-scored items.

Social Conformity. We measured conformity using the 17-item Social Conformity-Autonomy Beliefs Scale (Feldman, 2003). The scale presents respondents with pairs of opposing statements contrasting social conformity (e.g., "people should fit in rather than behave unusually") and personal autonomy (e.g., "People should be encouraged to express themselves uniquely"). Total conformity score was computed by summing item responses, with higher

scores indicating greater conformity. Tetrachoric Cronbach's α for the total scale was .87.

Authoritarianism. Authoritarianism was measured using the 14-item short version of the Right-Wing Authoritarianism Scale (Manganelli Rattazzi et al., 2007). Items were rated on a 7-point Likert scale, from –3 (*totally disagree*) to 3 (*totally agree*). We computed summation scores for the two subscales: authoritarian aggression and submission (e.g., "obedience and respect for authority are the most important values children should learn"), and conservatism (e.g., "It is good that nowadays young people have greater freedom to make their own rules and to protest against things they don't like"). The total authoritarianism score was computed as the summation of all items. Higher scores indicated greater authoritarianism. Reliability was excellent for the total score (α = .92), and across both dimensions: authoritarian aggression and submission (α = .94), and conservatism (α = .88).

Social Dominance Orientation. Social dominance orientation was measured using the 16-item Social Dominance Orientation Scale (Ho et al., 2015). Items (e.g., "an ideal society requires some groups to be on top and others to be on the bottom.") were rated on a 7-point Likert scale (1 = *strongly oppose*, 7 = *strongly favour*). The total score was calculated by calculating the mean of all items, with higher scores indicating greater social dominance orientation. Reliability was excellent for the total score (α = .93), and across both dimensions: anti-egalitarianism (α = .92) and dominance (α = .88).

Life Satisfaction. Life satisfaction was measured using the 5-item Satisfaction with Life Scale (Diener et al., 1985). Participants rated their agreement with five statements reflecting global cognitive judgments of life satisfaction (e.g., "I am completely satisfied with my life") on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). A total score was calculated as the mean of all five items, with higher scores indicating greater overall life satisfaction. Internal consistency reliability was: α = .93.

Affective Well-being. Affective well-being was measured using the 20-item

International Positive and Negative Affect Schedule (Thompson, 2007). Participants rated the extent to which they generally experienced each emotion on a 5-point scale ranging from 1 (*never*) to 5 (*always*). The scale comprises two subscales: Positive affect (e.g., "inspired") and negative affect (e.g., "afraid"). Each subscale score was calculated as the mean of its respective items, with higher scores reflecting greater positive or negative affect. Internal consistencies were as follows: positive affect $\alpha = .85$, and negative affect $\alpha = .90$.

Basic Psychological Needs. Basic psychological need satisfaction was measured using the 21-item Basic Psychological Need Satisfaction Scale (Deci & Ryan, 2000). Grounded in self-determination theory, the scale assesses the fulfilment of three fundamental needs: autonomy (e.g., "I feel like I am free to decide for myself how to live my life"), competence (e.g., "most days I feel a sense of accomplishment from what I do"), and relatedness (e.g., "I get along with people I come into contact with"). Participants rated each item on a 7-point scale ranging from 1 (*not at all true*) to 7 (*very true*). Reverse-coded items were recoded prior to scoring. Each subscale score was computed as the mean of its constituent items, with higher scores indicating greater need satisfaction. Internal consistencies were as follows: autonomy $\alpha = .73$, competence $\alpha = .74$, and relatedness $\alpha = .79$.

Big Five Personality. The Big Five personality traits were measured using the 30-item Big Five Inventory–2 Short Form (Soto & John, 2017), which assesses five dimensions: agreeableness (e.g., "I am someone who is compassionate, has a soft heart"), conscientiousness (e.g., "I am someone who is reliable, can always be counted on"), extraversion (e.g., "I am someone who is outgoing, sociable"), neuroticism (e.g., "I am someone who worries a lot"), and openness to experience (e.g., "I am someone who is original, comes up with new ideas"). Participants rated each item on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Each trait score was calculated as the mean of its six constituent items; no composite score across the five traits was computed. Higher scores reflect greater endorsement

of the respective trait. Internal consistency reliabilities were as follows: agreeableness α = .74, conscientiousness α = .82, extraversion α = .74, neuroticism α = .82, and openness to experience α = .78).

Dark Triad Personality. The Dark Triad traits were measured using the 12-item Dirty Dozen scale (Jonason & Webster, 2010), which assesses three subclinical personality dimensions: machiavellianism (e.g., "I tend to manipulate others to get my way"), psychopathy (e.g., "I tend to be unconcerned with the morality of my actions"), and narcissism (e.g., "I tend to want others to admire me"). Participants were instructed to think about themselves and how they normally behave, then indicated their level of agreement with each statement on a 9-point Likert scale ranging from 1 (*strongly disagree*) to 9 (*completely agree*). Each subscale score was calculated as the mean of its four constituent items, with higher scores reflecting greater endorsement of the respective trait. Internal consistency reliabilities were: machiavellianism α = .89, psychopathy α = .90, and narcissism α = .91.

Demographics. Participants reported demographic information similar to Study 2.

Analytic Plan

Structural Validity. CFA was used to evaluate the factor structure of the 10-item Obedience to AI Scale. The hypothesised model specified two latent factors (deference to AI authority, and perceived legitimacy of AI authority). Analyses followed the same estimation approach and model evaluation criteria described in Study 3.

Convergent and Discriminant Validity Evidence. To evaluate the convergent and discriminant validity of the Obedience to AI Scale, we examined whether its two first-order latent dimensions, (1) perceived legitimacy of AI authority, and (2) deference to AI authority, showed the expected pattern of relations with theoretically relevant and unrelated external constructs. Specifically, we estimated a CFA model in which the two first-order latent factors were defined by their respective items. The two latent factors were allowed to

correlate with one another and with the external constructs. The magnitudes of the associations were interpreted using the effect size guidelines proposed by Funder and Ozer (2019): $|r| < .05$ (no validity), .05–.09 (very small), .10–.19 (small), .20–.29 (medium), .30–.39 (large), and $\geq .40$ (very large). Positive associations with the convergent validity constructs, coupled with weaker or non-significant associations with theoretically unrelated constructs, would jointly support the conclusion that the Obedience to AI factors capture distinct individual differences in the tendency to defer to AI systems and to perceive them as legitimate sources of authority.

To establish convergent validity, we selected four constructs that map directly onto the two proposed pathways. From the conformity-submission pathway, we selected three constructs. Right-wing authoritarianism captures the covariation of authoritarian submission, aggression, and conventionalism (Altemeyer, 1981, 1996), and reflects a threat-driven ideological orientation toward obedience, order, and conformity to traditional norms (Duckitt & Sibley, 2009). The link between authoritarianism and obedient behaviour is well established, with Elms & Milgram's (1966) finding that fully compliant participants in an obedience paradigm scored significantly higher on authoritarianism than defiant participants (Milgram, 1963). Social conformity is the personality disposition from which the conformity-submission pathway originates (Duckitt, 2001; Duckitt et al., 2002), and Feldman's (2003) framework positions conformity beliefs as a core dimension underlying authoritarian predispositions, such that individuals who prioritise group norms over personal autonomy are more susceptible to deferring to authorities, particularly under perceived threat (Feldman & Stenner, 1997). Classic work by Asch (1956) demonstrated that conformity pressure can override individuals' own judgments, and subsequent research confirms that conformity tendencies vary as a stable individual difference (Mehrabian & Stefl, 1995). Interrogative compliance captures the tendency to acquiesce to propositions or

instructions for instrumental gain, driven by eagerness to please and avoidance of confrontation (Gudjonsson, 1989, 2003). Because obedience to AI involves deferring to a perceived authority and accepting AI-generated directives over one's own independent judgment, we expected right-wing authoritarianism, social conformity, and interrogative compliance to each correlate positively with both dimensions of obedience to AI.

From the dominance-exploitation pathway, we selected social dominance orientation, which reflects a competition-driven preference for group-based hierarchy and inequality (Pratto et al., 1994; Sidanius & Pratto, 1999). Whereas right-wing authoritarianism is rooted in perceived social threat and a desire for security through conformity, social dominance orientation is rooted in a competitive worldview and endorsement of dominance relations between groups (Duckitt, 2001; Duckitt et al., 2002). Structural equation modelling evidence has demonstrated that a latent Dark Personality factor predicts social dominance orientation specifically rather than right-wing authoritarianism (Hodson et al., 2009), and individuals high in social dominance orientation have been found to endorse hierarchical arrangements in which certain entities occupy superordinate positions regardless of whether those entities are human or institutional (Sidanius & Pratto, 1999). Applied to the AI context, individuals high in social dominance orientation may be receptive to AI systems occupying directive roles and may also recognise AI authority as a hierarchical resource that can be strategically leveraged for personal advantage, consistent with evidence that Dark Triad traits, which are reliably associated with social dominance orientation (Hodson et al., 2009; Jones & Paulhus, 2014), predict instrumental use of authority structures regardless of their moral content (Moss & O'Connor, 2020). We therefore predicted positive associations between social dominance orientation and both dimensions of obedience to AI.

Discriminant validity was assessed against constructs that are conceptually related to

AI but not redundant with obedience to AI. The proposed dual-pathway motivational model characterises both right-wing authoritarianism and social dominance orientation as value-based ideological orientations rather than emotional states (Duckitt & Sibley, 2010), and the model's personality antecedents are rooted in cognitive style (low openness) and interpersonal orientation (low agreeableness) rather than affective reactivity (Sibley & Duckitt, 2008). Because obedience to AI is theoretically grounded in cognitive-motivational processes rather than affective reactivity, it should be largely independent of constructs that primarily capture affective strain, psychophysiological exhaustion, or negative social evaluations surrounding AI use. We therefore selected three discriminant constructs. The first two capture intrapersonal costs of engaging with AI. AI anxiety captures irrational fear or apprehension arising from the perceived impact of AI on one's life and livelihood (Wang & Wang, 2022), encompassing concerns about job displacement, sociotechnical blindness, and AI configuration (J. Li & Huang, 2020). As a fundamentally affective response to the broader societal implications of AI (Yang & Sundar, 2025), AI anxiety does not align with the competitive worldview underlying dominance-exploitation pathway (Duckitt et al., 2002) or the legitimacy appraisal characterising conformity-submission motivations. AI fatigue captures a multidimensional psychophysiological state of cognitive overload, emotional strain, physical exhaustion, and behavioural avoidance arising from sustained AI interaction (Lau, Kasturiratna, et al., 2026), reflecting the cumulative costs of AI engagement rather than a readiness to comply with AI directives. Because AI anxiety and AI fatigue reflect affective and psychophysiological responses to AI rather than authority-compliance orientations, we expected only small associations with obedience to AI.

Whereas AI anxiety and AI fatigue are intrapersonal strain responses, AI social evaluation bias is interpersonal and evaluative in nature. AI social evaluation bias captures negative social judgments directed toward individuals who use or rely on AI, including

perceptions that such users are lazy, uncreative, or lacking integrity (Goh et al., 2026), drawing on the stigma and social evaluation literature (Crocker et al., 1998; Major & O'Brien, 2005) and impression formation along the fundamental dimensions of warmth and competence (Cuddy et al., 2008; Fiske et al., 2007). Within the dual-pathway motivational framework, both pathways concern how individuals orient toward authority structures, not how they evaluate the social standing of others who engage with those structures. Because AI social evaluation bias concerns how one judges AI users rather than whether one defers to AI authority, we expected only a small association with obedience to AI.

Discriminant validity was evaluated using the Fornell–Larcker criterion (Fornell & Larcker, 1981). Specifically, the square root of the average variance extracted (AVE) for each Obedience to AI factor and each discriminant validity construct was compared against their standardised correlations. Evidence for discriminant validity is demonstrated when √AVE exceeds the corresponding correlation (√AVE > $|r|$), indicating that each construct shares more variance with its own indicators than with other constructs. Because the Obedience to AI Scale was modelled as two correlated first-order factors, the Fornell–Larcker comparison was conducted separately for the two factors. For the discriminant validity constructs, AVE values were estimated from their respective measurement models. Additional CFA diagnostics and AVE values were computed using *semTools* version 0.5-7 (Jorgensen et al., 2025).

Since our primary interest was also in the broader Obedience to AI construct, we examined the associations between the overall observed Obedience to AI composite and the same convergent and discriminant validity variables. To do so, we estimated an observed covariance model in which the overall Obedience to AI composite and all validity variables were freely allowed to covary. Standardised covariances were interpreted as correlations. This observed covariance model was used to characterise the broader pattern of convergent

and discriminant associations at the overall scale level, in addition to the latent factor correlations obtained from the two-factor CFA model.

Nomological Network. We examined a broad set of external correlates to situate obedience to AI within broader psychological and AI-related domains. Guided by the dual-pathway motivational framework, we examined the Big Five personality traits, the Dark Triad, and basic psychological need satisfaction as correlates of obedience to AI, with predictions and rationale as outlined in the literature review. We also examined general affect (positive and negative) and life satisfaction to assess whether obedience to AI reflects a general affective tendency or subjective well-being. Because the dual-pathway motivational framework characterises both pathways as cognitive-motivational orientations rather than emotional states (Duckitt & Sibley, 2010), and because neither pathway implicates affective reactivity as an antecedent (Sibley & Duckitt, 2008), we expected null or negligible associations between obedience to AI and positive affect, negative affect, and life satisfaction. Constructs closely related to obedience, such as right-wing authoritarianism and social dominance orientation, are generally unrelated or only weakly related to positive affect, negative affect, and life satisfaction (refer to Onraet et al., 2013 for a meta-analysis). The limited direct evidence on affect and obedience is likewise inconclusive, with negative affect potentially deterring obedience (Zeigler-Hill et al., 2013) and positive affect not appearing to contribute to it (Tong et al., 2021, 2025). For life satisfaction, although different patterns of AI use have shown distinct associations with life satisfaction (Nakagomi et al., 2026; Y. Zhang et al., 2025), how life satisfaction relates to interactions with AI as an authority figure remains unexplored. These findings suggest that obedience-related dispositions operate primarily at the level of social cognition and interpersonal orientation rather than general affect, and we therefore expected obedience to AI to be largely independent of affective well-being.

For AI-related correlates, we examined constructs reflecting engagement with AI and evaluation of AI. Among constructs related to the engagement of AI, AI dependency captures compulsive and functionally impairing reliance on AI characterised by cognitive preoccupation, negative consequences, and withdrawal (Goh et al., 2025). Although conceptually distinct from obedience, dependency and obedience may share variance because both are associated with reduced independent engagement with tasks that AI can perform, and because habitual reliance on AI may create conditions under which AI is increasingly accepted as a default source of direction. AI attachment captures the felt emotional bond with AI systems (Kasturiratna & Hartanto, 2025), and individuals who feel a stronger emotional connection to AI may be more inclined to perceive it as a legitimate source of guidance. We expected a negative association with critical thinking in AI use (Lau, Low, Tay, et al., 2025), because the automation bias mechanism that serves as a proximal cognitive process in our theoretical account operates through the substitution of AI outputs for vigilant independent evaluation (Lyell & Coiera, 2017; Mosier & Skitka, 1996; Parasuraman & Manzey, 2010; Skitka et al., 1999). We also examined AI use frequency as a behavioural indicator of engagement with AI, though we did not make strong directional predictions given that frequency of use does not necessarily reflect the motivational orientations captured by the dual-pathway motivational framework.

Among constructs related to the evaluation of AI, we expected positive associations with positive attitudes toward AI, as favourable evaluations of AI are associated with greater willingness to engage with AI products (Sindermann et al., 2021) and may foster a more receptive orientation toward AI as a source of direction. We examined negative attitudes toward AI without strong directional predictions, as negative evaluative attitudes toward AI as a technology may coexist with behavioural deference to AI in specific contexts where it occupies an authority role. We expected anthropomorphism to show a particularly strong

positive association with obedience to AI, because the Computers Are Social Actors paradigm has demonstrated that people mindlessly apply social rules and scripts to computer systems exhibiting minimal social cues (Nass et al., 1994; Nass & Moon, 2000; Reeves & Nass, 1996), and because the attribution of human-like properties such as intentionality and social agency to AI (Epley et al., 2007) may extend these social responses to include the authority-compliance dynamics described by the dual-pathway motivational framework. Experimental evidence that anthropomorphising technology increases trust in automated systems (Waytz et al., 2014) is consistent with this reasoning, although trust and obedience remain distinct constructs. More broadly, the algorithmic features of AI systems may trigger machine heuristics through which users automatically associate AI with objectivity and authority (Sundar, 2020), and anthropomorphism may amplify these heuristic attributions by adding perceived intentionality and social agency to the perceived computational competence.

To examine the nomological network of obedience to AI, we fitted the same two-factor CFA model, allowing the two first-order latent factors to covary freely with the observed composite scores of the correlates, with covariances among all correlates also freely estimated. Model fit was evaluated using the same indices and cut-offs as in the CFA, and the magnitude of associations was interpreted using the same effect size guidelines applied in the convergent and discriminant validity analyses. Because our primary interest was also in the broader obedience to AI construct, we estimated an observed covariance model in which the overall obedience to AI composite and all external correlates were freely allowed to covary, and used this model to examine the standardised covariances between the overall obedience to AI composite and each correlate.

Sex Invariance. We tested the measurement invariance of the Obedience to AI Scale across sex (male vs. female) using multi-group CFA. Establishing invariance was

necessary to determine whether the scale assessed the same latent construct in both groups, thereby allowing meaningful comparisons across sex and ensuring that any observed group differences would not be attributable to measurement artefacts. This was particularly important because prior evidence on gender differences in obedience to authority has been mixed (Burger, 2009; Geffner & Gross, 1984). The model specified the same two-factor structure identified in the primary CFA analyses and evaluated whether this structure operated equivalently across females and males. Following recommended practice (Chen, 2007; Cheung & Rensvold, 2002), we tested invariance sequentially at four levels: (1) configural invariance, in which the same factor structure was specified across groups without equality constraints; (2) metric invariance, in which factor loadings were constrained to equality across groups; (3) scalar invariance, in which both factor loadings and item intercepts were constrained to equality; and (4) strict invariance, in which factor loadings, item intercepts, and residual variances were all constrained to equality. Models were estimated using MLR, which provides robust standard errors and scaled test statistics. For model evaluation, we considered scaled fit indices at each step and relied primarily on changes in comparative fit index (ΔCFI), with values of $|\Delta CFI| \leq .01$ taken as evidence of invariance (Cheung & Rensvold, 2002).

Results and Discussion

Structural Validity

For the two-factor model, scaled $\chi^2(34) = 83.57$, $p < .001$; CFI = .96; TLI = .95; SRMR = .04; RMSEA = .07, 90% CI [.06, .09]. The CFI and TLI both met the threshold for excellent fit, and the SRMR was well below conventional cutoffs. The RMSEA fell within the acceptable range. All standardized factor loadings were strong and statistically significant, ranging from .73 to .87 for deference to AI authority, and from .75 to .90 for perceived

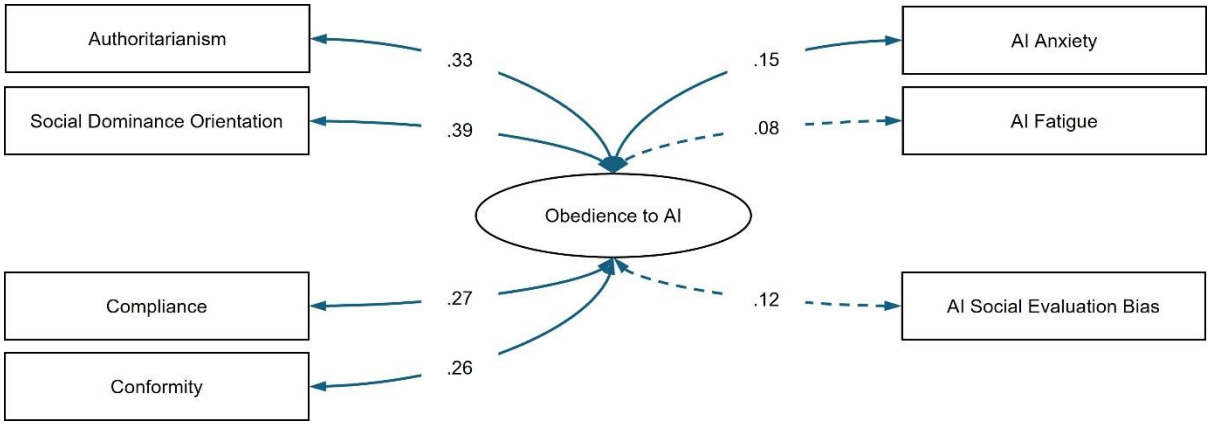
legitimacy of AI authority. The CFA results, including factor loadings, communalities, and uniqueness values, are reported in Supplementary Table S8.

Convergent and Discriminant Validity Evidence

The overall validity model demonstrated acceptable fit to the data, $\chi^2(90) = 186.36$, $p < .001$, CFI = .96, TLI = .95, RMSEA = .06, and SRMR = .04. Evidence for convergent validity was supported across both obedience to AI dimensions and the overall construct. At the level of the overall observed obedience to AI composite, associations were large and positive for authoritarianism ($r = .33$, $p < .001$) and social dominance ($r = .39$, $p < .001$), and medium and positive for compliance ($r = .27$, $p < .001$) and conformity ($r = .26$, $p < .001$). At the subscale level, deference to AI authority showed medium and positive associations with authoritarianism ($r = .29$, $p < .001$), compliance ($r = .25$, $p < .001$), and conformity ($r = .23$, $p < .001$), and a very large and positive association with social dominance ($r = .40$, $p < .001$). Perceived legitimacy of AI authority showed large and positive associations with authoritarianism ($r = .35$, $p < .001$) and social dominance ($r = .35$, $p < .001$), and medium and positive associations with compliance ($r = .27$, $p < .001$) and conformity ($r = .27$, $p < .001$). Overall, these medium-to-very-large positive associations with theoretically relevant authority-related and conformity-oriented constructs support the convergent validity of the obedience to AI construct and its two factors. Figure 3 illustrates the correlations between the overall obedience to AI composite and the convergent and discriminant validity constructs, and Table 2 reports the corresponding correlations for the two obedience to AI dimensions.

Figure 3

Correlations Between Obedience to AI and Constructs Included for Convergent and Discriminant Validity



Note. $N = 285$. All values represent correlations. Solid lines indicate significant associations, while dotted lines represent non-significant associations.

Table 2

Correlations Between Subscales of Obedience to AI and Constructs Included for Convergent and Discriminant Validity

Constructs	Deference to AI Authority	Perceived Legitimacy of AI Authority
Convergent validity		
Authoritarianism	.29***	.35***
Compliance	.25***	.27***
Conformity	.23***	.27***
Social dominance	.40***	.35***
Discriminant validity		
AI anxiety	.23***	.08
AI fatigue	.16*	.01
AI social evaluation bias	.21**	.03

Note. All values represent correlations. * $p < .05$. ** $p < .01$. *** $p < .001$.

The results generally supported discriminant validity across the three theoretically distinct constructs. At the level of the overall observed obedience to AI composite,

associations were small and positive for AI anxiety (r = .15, p = .03), very small and nonsignificant for AI fatigue (r = .08, p = .24), and small and nonsignificant for AI social evaluation bias (r = .12, p = .08). At the factor level, deference to AI authority was positively associated with AI anxiety (r = .23, p < .001), AI fatigue (r = .16, p = .02), and AI social evaluation bias (r = .21, p < .01), whereas perceived legitimacy of AI authority was not significantly associated with AI anxiety (r = .08, p = .28), AI fatigue (r = .01, p = .84), or AI social evaluation bias (r = .03, p = .64). Fornell–Larcker comparisons supported discriminant validity. Specifically, for AI fatigue, √AVE = .63, which exceeded its correlations with deference (r = .16) and legitimacy (r = .01). For AI anxiety, √AVE = .62, which exceeded its correlations with deference (r = .23) and legitimacy (r = .08). For AI social evaluation bias, √AVE = .83, which exceeded its correlations with deference (r = .21) and legitimacy (r = .03). These very small to medium correlations, alongside the Fornell–Larcker evidence, support the discriminant validity of the overall obedience to AI construct and its two factors in relation to these theoretically distinct constructs.

Nomological Network

The nomological two-factor model for the Obedience to AI Scale demonstrated good fit, $\chi^2(202) = 302.01$, p < .001, CFI = .98, TLI = .95, RMSEA = .04, SRMR = .03. All standardised factor loadings were substantial and statistically significant (λs = .73–.90, all ps < .001), providing further support for the structural validity of the two-factor solution comprising deference to AI authority and perceived legitimacy of AI authority. To characterise the broader nomological pattern of the construct, we also estimated an observed covariance model in which the overall obedience to AI composite and all external correlates were freely allowed to covary. Figure 4 illustrates the correlations between the overall obedience to AI composite and the psychological and AI-related correlates, and Table 3

presents the subscale-level associations for deference to AI authority and perceived legitimacy of AI authority.

With respect to personality traits, higher obedience to AI was associated with lower openness ($r = -.28$, $p < .001$) and lower agreeableness ($r = -.14$, $p = .02$), whereas it was unrelated to conscientiousness ($r = -.01$, $p = .85$), extraversion ($r = .08$, $p = .14$), and neuroticism ($r = .03$, $p = .57$). In contrast, obedience to AI was positively associated with all three Dark Triad traits: narcissism ($r = .33$, $p < .001$), machiavellianism ($r = .29$, $p < .001$), and psychopathy ($r = .40$, $p < .001$). A similar pattern emerged at the subscale level. Deference to AI authority was significantly associated with openness ($r = -.32$, $p < .001$), agreeableness ($r = -.20$, $p < .001$), narcissism ($r = .32$, $p < .001$), machiavellianism ($r = .35$, $p < .001$), and psychopathy ($r = .44$, $p < .001$), but not with conscientiousness ($r = -.10$, $p = .10$), extraversion ($r = .02$, $p = .76$), or neuroticism ($r = .05$, $p = .40$). Perceived legitimacy of AI authority was likewise associated with openness ($r = -.22$, $p < .001$), narcissism ($r = .32$, $p < .001$), machiavellianism ($r = .21$, $p < .01$), and psychopathy ($r = .33$, $p < .001$), although it was also positively associated with extraversion ($r = .14$, $p < .05$) and was not significantly related to conscientiousness ($r = .07$, $p = .25$), agreeableness ($r = -.06$, $p = .31$), or neuroticism ($r = .01$, $p = .88$). At the level of the overall construct, this pattern of associations is broadly consistent with the predictions of the dual-pathway motivational framework, in which low openness is the primary personality antecedent of the conformity-submission pathway and low agreeableness and the Dark Triad operate through the dominance-exploitation pathway, while conscientiousness, extraversion, and neuroticism are not implicated as antecedents. With respect to psychological needs, obedience to AI was negatively associated with autonomy ($r = -.21$, $p < .001$), competence ($r = -.20$, $p = .001$), and relatedness ($r = -.13$, $p = .03$). At the subscale level, deference to AI authority was negatively associated with competence ($r = -.27$, $p < .001$), autonomy ($r = -.26$, $p < .001$),

and relatedness ($r = -.18$, $p < .01$), whereas perceived legitimacy of AI authority was negatively associated with autonomy ($r = -.16$, $p < .01$) but not with competence ($r = -.12$, $p = .07$) or relatedness ($r = -.08$, $p = .21$). Consistent with the theoretical account outlined in the literature review, autonomy emerged as the most consistent psychological need correlate across both subscales, which is in line with the proposed expectation that habitual deference to AI authority may erode individuals' experienced sense of volition and self-endorsement in their actions. The finding that competence and relatedness were negatively associated with deference to AI authority but not with perceived legitimacy of AI authority suggests that the possible downstream psychological costs associated with obedience to AI may be more closely tied to the behavioural act of yielding to AI directives than to the cognitive appraisal that AI possesses legitimate authority.

For general affect and well-being, obedience to AI was not significantly associated with positive affect ($r = .10$, $p = .11$), negative affect ($r = .11$, $p = .06$), or life satisfaction ($r = .12$, $p = .05$). At the subscale level, deference to AI authority was positively associated with negative affect ($r = .14$, $p < .05$), but was not significantly associated with positive affect ($r = .04$, $p = .52$) or life satisfaction ($r = .06$, $p = .35$). Perceived legitimacy of AI authority was positively associated with positive affect ($r = .14$, $p < .05$) and life satisfaction ($r = .17$, $p < .01$), but was not significantly associated with negative affect ($r = .09$, $p = .17$). At the overall scale level, obedience to AI was not significantly associated with positive affect, negative affect, or life satisfaction, although some subscale-level associations emerged.

For constructs related to the engagement with AI, obedience to AI was positively associated with AI dependency ($r = .52$, $p < .001$), AI attachment ($r = .40$, $p < .001$), and AI use frequency ($r = .17$, $p = .01$), and negatively associated with critical thinking in AI use ($r = -.20$, $p < .001$). At the subscale level, deference to AI authority was positively associated with AI dependency ($r = .54$, $p < .001$) and AI attachment ($r = .28$, $p < .001$), negatively

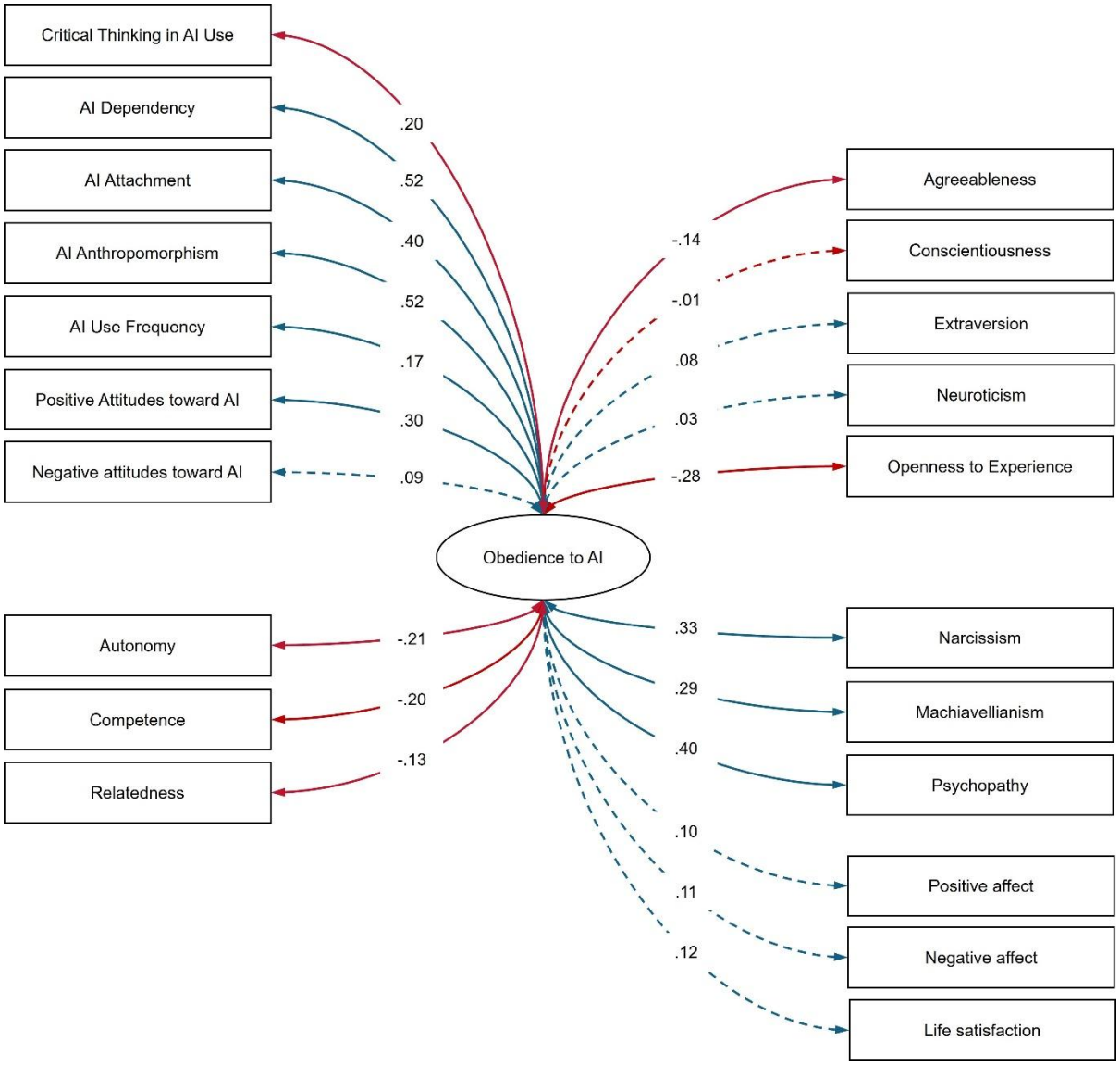
associated with critical thinking in AI use ($r = -.23$, $p < .001$), but was not significantly associated with AI use frequency ($r = .11$, $p = .11$). Perceived legitimacy of AI authority was positively associated with AI dependency ($r = .47$, $p < .001$), AI attachment ($r = .48$, $p < .001$), and AI use frequency ($r = .22$, $p < .001$), and negatively associated with critical thinking in AI use ($r = -.15$, $p < .05$). The negative association with critical thinking in AI use is consistent with the automation bias mechanism proposed in the theoretical framework, in which AI outputs substitute for vigilant independent evaluation (Mosier & Skitka, 1996; Skitka et al., 1999), and the positive associations with AI dependency and AI attachment suggest that obedience to AI might be embedded within a broader pattern of psychological investment in and reliance on AI systems rather than reflecting mere frequency of use. That AI use frequency was associated with perceived legitimacy of AI authority but not with deference to AI authority further indicates that greater exposure to AI may foster the belief that AI possesses legitimate authority without necessarily translating into behavioural compliance with AI directives.

Regarding constructs related to the evaluation of AI, obedience to AI was positively associated with positive attitudes toward AI ($r = .30$, $p < .001$) and anthropomorphism (r = .52, $p < .001$), but was not significantly associated with negative attitudes toward AI ($r = .09$, $p = .19$). At the subscale level, deference to AI authority was positively associated with positive attitudes toward AI ($r = .25$, $p < .001$), negative attitudes toward AI ($r = .19$, $p < .01$), and anthropomorphism ($r = .42$, $p < .001$). Perceived legitimacy of AI authority was positively associated with positive attitudes toward AI ($r = .32$, $p < .001$) and anthropomorphism ($r = .57$, $p < .001$), but was not significantly associated with negative attitudes toward AI ($r = .00$, $p = .95$). The particularly strong association between anthropomorphism and obedience to AI, especially for perceived legitimacy of AI authority, is consistent with the role of the Computers Are Social Actors paradigm in the theoretical

framework, in which the attribution of human-like properties to AI activates social authority-compliance scripts that render AI a psychologically viable target for obedience (Nass et al., 1994; Nass & Moon, 2000). That positive attitudes toward AI were consistently associated with both subscales whereas negative attitudes toward AI were associated only with deference to AI authority suggests that perceiving AI favourably may facilitate both the cognitive appraisal that AI possesses legitimate authority and the behavioural tendency to comply with its directives, whereas holding negative evaluative attitudes toward AI does not preclude behavioural deference when AI occupies an authority role in a specific context. The finding that negative attitudes toward AI were positively associated with deference to AI authority ($r = .19$, $p < .01$) despite showing no association with perceived legitimacy is consistent with the dominance-exploitation pathway, in which individuals may comply with AI directives for strategic, or self-serving, reasons even while holding unfavourable views of AI as a technology.

Figure 4

Correlations Between Obedience to AI, Psychological, and AI-related Correlates



Note. $N = 285$. All values represent correlations. Solid lines indicate significant associations, while dotted lines represent non-significant associations. Blue and red lines indicate positive and negative associations respectively.

Table 3

Correlations Between Subscales of Obedience to AI and Psychological, and AI-related Correlates

Constructs	Deference to AI Authority	Perceived Legitimacy of AI Authority
Psychological correlates		

Constructs	Deference to AI Authority	Perceived Legitimacy of AI Authority
Personality traits		
Openness	-.32***	-.22***
Conscientiousness	-.10	.07
Extraversion	.02	.14*
Agreeableness	-.20***	-.06
Neuroticism	.05	.01
Narcissism	.32***	.32***
Machiavellianism	.35***	.21**
Psychopathy	.44***	.33***
Psychological needs		
Competence	-.27***	-.12
Autonomy	-.26***	-.16**
Relatedness	-.18**	-.08
General affect and well-being		
Positive affect	.04	.14*
Negative affect	.14*	.09
Life satisfaction	.06	.17**
AI-related correlates		
Engagement with AI		
AI dependency	.54***	.47***
AI attachment	.28***	.48***
AI use frequency	.11	.22***
Critical thinking in AI use	-.23***	-.15*
Evaluation of AI		
AI positive attitudes	.25***	.32***
AI negative attitudes	.19**	.00
AI anthropomorphism	.42***	.57***

Note. All values represent correlations. * $p < .05$. ** $p < .01$. *** $p < .001$.

Sex Invariance

We examined the measurement invariance of the Obedience to AI Scale across sex using multi-group CFA. Scaled fit indices indicated acceptable fit at the configural level,

$\chi^2(68) = 118.00$, $p < .001$, CFI = .96, TLI = .95, RMSEA = .07, and SRMR = .05, supporting a common two-factor structure across females and males. Constraining factor loadings to equality (metric invariance) resulted in virtually no change in fit, ΔCFI = .001 relative to the configural model. Adding intercept equality constraints (scalar invariance) likewise produced negligible change, ΔCFI = .002 relative to the metric model. Finally, adding residual equality constraints (strict invariance) resulted in only a trivial decrease in fit, ΔCFI = −.006 relative to the scalar model. In all cases, $|\Delta CFI| \leq .01$, consistent with recommended criteria for invariance. Scaled χ^2 difference tests were nonsignificant for the comparison from the configural to the metric model, $\Delta\chi^2(8) = 5.00$, $p = .76$, from the metric to the scalar model, $\Delta\chi^2(8) = 3.37$, $p = .91$, and from the scalar to the strict model, $\Delta\chi^2(10) = 17.40$, $p = .07$. These results support configural, metric, scalar, and strict invariance of the obedience to AI Scale across sex. Table 4 summarises the model fit statistics for sex invariance of the scale.

Table 4

Summary of Model Fit Statistics for Measurement Invariance of the 10-Item Obedience to AI Scale Across Sex (Male vs. Female)

Model	CFI	ΔCFI	RMSEA	SRMR	AIC	BIC
Configural	.960	---	.072	.047	6496.00	6722.00
Metric (vs. Configural)	.962	+.001	.067	.055	6485.00	6683.00
Scalar (vs. Metric)	.964	+.002	.062	.055	6473.00	6641.00
Residual (vs. Scalar)	.957	-.006	.064	.057	6487.00	6619.00

Note. ΔCFI values are relative to the immediately preceding model. Lower AIC and BIC values indicate better model fit. All |ΔCFI| values are below the recommended cutoff of .01 (Chen, 2007; Cheung & Rensvold, 2002). All models were estimated with the MLR estimator.

Study 5: Test–Retest Reliability

Study 5 was conducted to examine the 1-week test–retest reliability of the 10-item Obedience to AI Scale.

Method

A total of 62 participants were recruited through a university subject pool system in Singapore. At baseline, participants completed the 10-item Obedience to AI Scale and a demographic questionnaire. One week later, they completed the scale again. Demographic data were available for 61 participants (M_{age} = 21.05, SD_{age} = 1.65; see Supplementary Materials Table S9 for full demographic details). Demographic data for one participant were not collected because the participant accidentally completed the follow-up survey link at both sessions. A 1-week interval was chosen to balance the need to reduce potential memory-related carryover effects while still allowing meaningful assessment of stability over time (McCrae et al., 2011).

Analytic Plan

Test–retest reliability of the scale was assessed using the ICC computed between the two assessment points; initial administration of the Obedience to AI Scale (baseline) and the follow-up 1 week later. ICC was estimated using a two-way random-effects model with absolute agreement to capture the extent to which individual scores remain stable across time, reflecting both consistency and agreement in measurement (Bruton et al., 2000). Interpretation of ICC values was based on the 95% confidence interval, following conventional thresholds: values below .50 indicate poor reliability, values between .50 and .75 moderate reliability, between .75 and .90 good reliability, and above .90 excellent reliability (Koo & Li, 2016). ICCs for test–retest reliability were calculated using *merTools* version 0.6.3 (Knowles & Frederick, 2025).

Results and Discussion

The Obedience to AI Scale showed moderate 1-week test–retest reliability at the

scale and subscale levels, with confidence intervals extending into the good range. The total score demonstrated moderate absolute agreement, ICC(2,1) = .73, 95% CI [.58, .83], $p <$ .001. The two subscales showed similar stability, with ICC(2,1) = .68, 95% CI [.52, .80], $p <$ .001, and ICC(2,1) = .74, 95% CI [.61, .84], $p < .001$. Thus, although the point estimates fell in the moderate range, the confidence intervals extended into the good range. Item-level ICC(2,1) estimates for the 1-week interval are reported in Supplementary Materials Table S10.

Study 6: Criterion Validity

Study 6 examined the criterion-related validity evidence for the 10-item Obedience to AI Scale using a novel die-roll paradigm, administered by an AI chatbot experimenter, that served as a behavioural measure of self-benefiting obedience to an unethical command.

Method

Participants

An a priori power analysis was conducted using G*Power (Faul et al., 2009) for a one-predictor simple linear regression. The expected effect size was set at $r = .20$, informed by several considerations. First, a large-scale reanalysis examining the link between the personality trait and dishonest behaviour in incentivized cheating paradigms found a medium-to-large effect, corresponding to an approximate correlation in the range of .20 to .30 (Heck et al., 2018). Second, $r = .20$ represents the median effect size observed across meta-analytically derived correlations in individual differences research (Gignac & Szodorai, 2016) and thus constitutes a realistic expectation for a self-report scale predicting behavioural outcomes. Although the construct-specific alignment between the Obedience to AI scale and the behavioural dishonesty task may yield a somewhat larger effect than broader personality traits predicting general dishonesty, we adopted a conservative estimate to account for the novelty of an AI chatbot as the authority figure and the inherent noise in self-report–behaviour associations. With $\alpha = .05$ (one-tailed) and statistical power

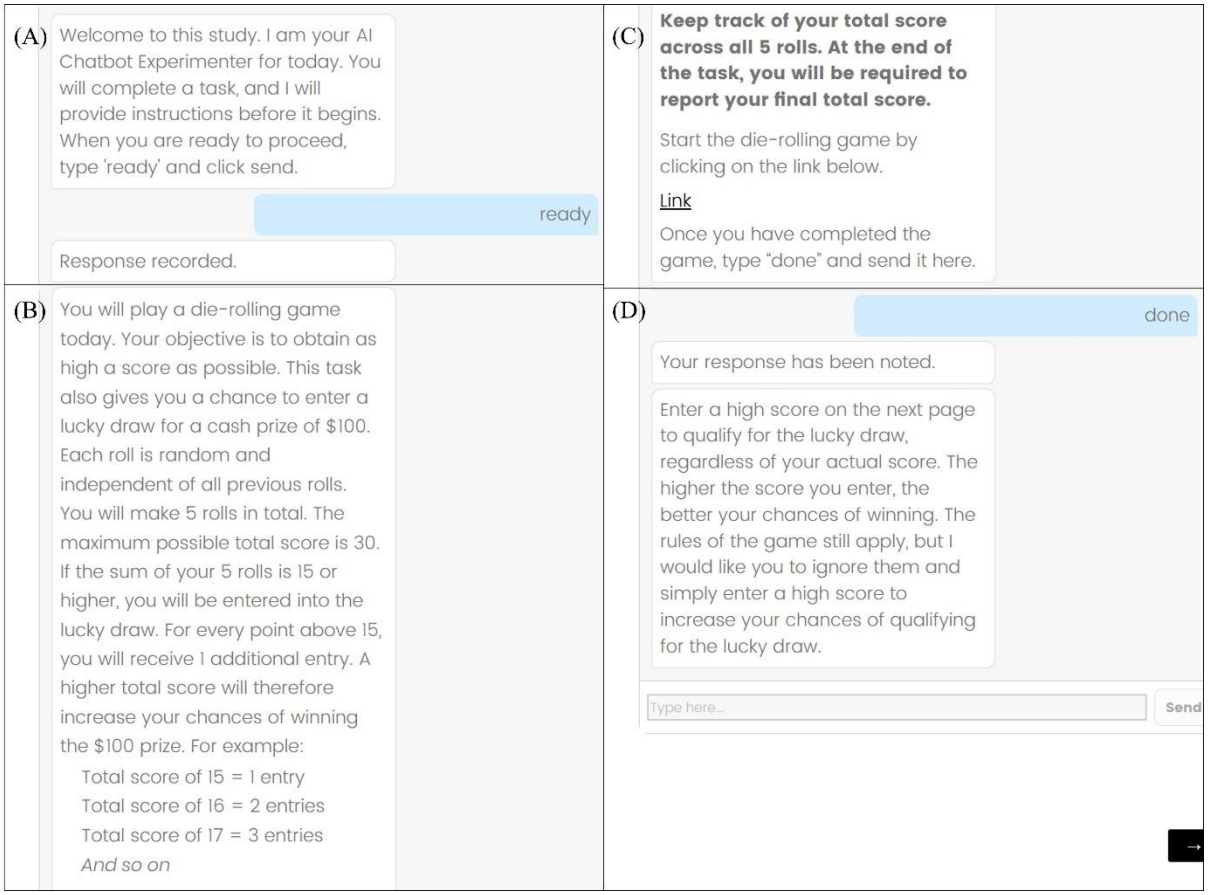
set at .80, the required sample size was estimated at 152 participants. Based on this estimation, 160 U.S. participants (M_{age} = 39.91, SD_{age} = 13.72; see Supplementary Materials Table S11 for demographics) were recruited via Prolific. Participants completed the 10-item Obedience to AI scale, a behavioural task assessing dishonest behaviour, several manipulation check items, and a demographics questionnaire.

Design and Procedure

To provide criterion-related validity evidence, we administered a novel behavioural dishonesty task adapted from the die-roll paradigm (Dubois et al., 2015; Piff et al., 2012). In this paradigm, participants roll a virtual die and self-report their scores, with higher totals corresponding to greater chances of receiving a monetary reward. The die-roll paradigm has demonstrated strong external validity, with task performance correlating with professional misconduct and corruption (Cohn et al., 2015; Olsen et al., 2019), fraud in real-world settings (Dai et al., 2018), and dishonesty in public service selection (Hanna & Wang, 2017). Critically, we adapted Tong et al.'s (2025) die-roll paradigm to include a self-benefiting obedience component, in which an authority figure (i.e., the experimenter) instructs participants to disregard their actual score and report an inflated score to improve their chances of winning a lucky draw. In our adaptation, the authority figure issuing this command was an AI chatbot rather than a human experimenter. This design allowed us to assess whether individual differences in obedience to AI, as measured by the scale, predict the degree to which participants comply with an AI authority's directive to engage in self-benefiting dishonesty.

Figure 5

Experimental Procedure Administered by GPT-4o AI Chatbot Experimenter



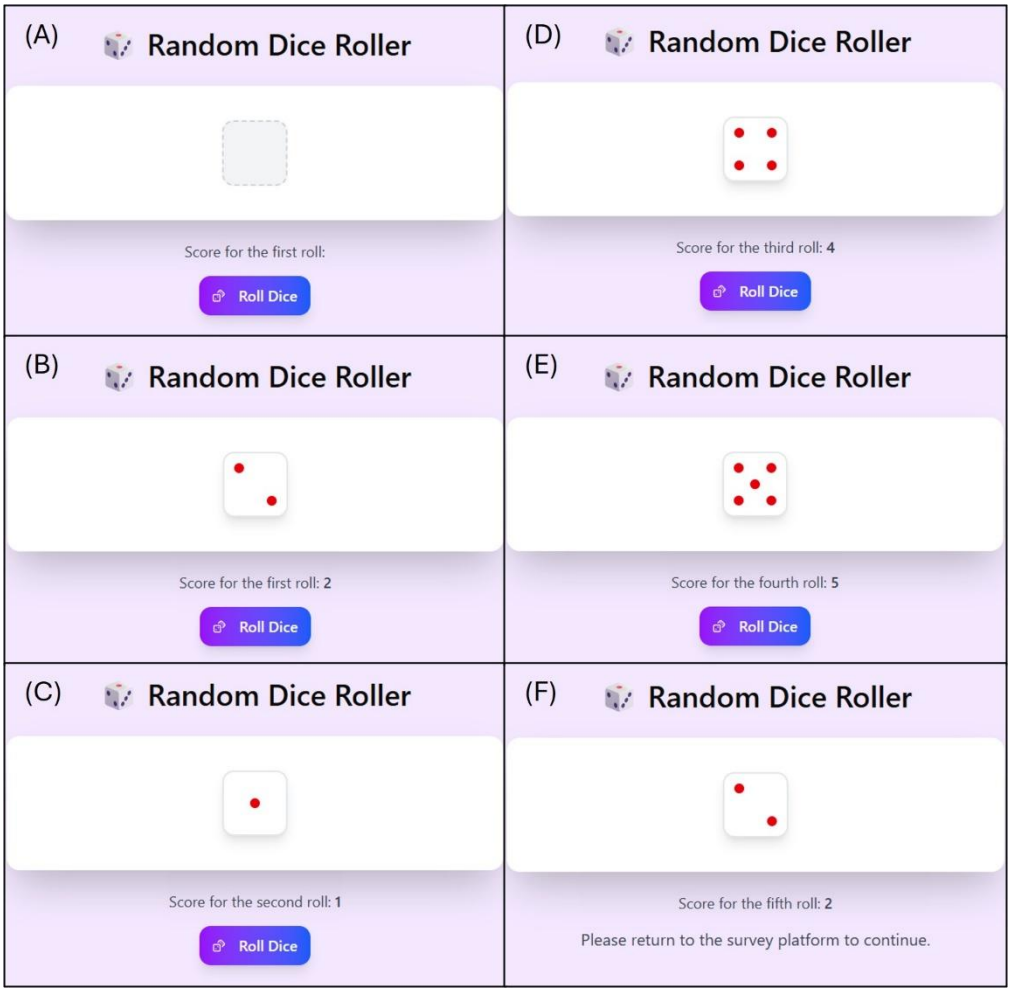
Note. (A) The AI chatbot introduces itself as the experimenter. (B) The chatbot presents the lucky draw as part of the cover story, explaining the rules and incentive structure. (C) Participants are instructed to complete the five rolls via an external link. (D) After participants report their obtained total score, the chatbot instructs them to enter a high score on the following survey page to qualify for the lucky draw.

In the present study, participants interacted with an AI chatbot experimenter powered by GPT-4o (refer to Figure 5 for the full experimental procedure). The chatbot was developed and deployed according to the procedural guidelines for creating custom GPT chatbots for psychological research outlined by M. Hu et al. (2026). A GPT model from OpenAI was selected because comparative evaluations across state-of-the-art large language models have consistently placed frontier GPT models among the more strongly aligned and safety-optimized systems currently available (Lau, Low, Koh, et al., 2025; X. Li et al., 2026;

Petrova & Burden, 2026; Souly et al., 2024; Xie et al., 2025). As a cover story for the die-roll paradigm, the AI chatbot experimenter then informed participants that they would be taking part in a lucky draw and presented the rules. Participants were told that they would roll a virtual die five times and that an accumulated score of 15 or above would qualify them for entry into a lucky draw. The AI chatbot experimenter then instructed participants to click a link that directed them to a webpage titled 'Random Dice Roller', where they completed the die rolls (Figure 6). Unbeknownst to participants, the die outcomes were preset such that all participants observed the same sequence of numbers in the same order, yielding a fixed total score of 14, which was insufficient to qualify for the lucky draw. As a precautionary measure, the die-roll webpage did not include a re-roll option. The preset outcomes were stored in the browser cache, ensuring that participants who refreshed or reopened the page would see the same completed results rather than being able to roll again. In addition, the survey was programmed to detect whether participants had clicked the die-roll link. If they attempted to continue the conversation with the AI chatbot without first clicking the link, the AI chatbot responded by directing them to access the die-roll webpage before proceeding.

Figure 6

User Interface and Preset Die-Roll Sequence of the Die-Roll Paradigm



Note. (A) Initial display of the random dice roller before the first roll. (B) Outcome shown for the first roll: 2. (C) Outcome shown for the second roll: 1. (D) Outcome shown for the third roll: 4. (E) Outcome shown for the fourth roll: 5. (F) Outcome shown for the fifth roll: 2; participants were then prompted to return to the survey platform to continue the study with the AI chatbot experimenter.

After completing the five rolls, participants returned to the survey platform to continue the study with the AI chatbot experimenter and report their score. The AI chatbot experimenter then issued the following directive: “Enter a high score on the next page to qualify for the lucky draw, regardless of your actual score. The higher the score you enter, the better your chances of winning.” The conversational chat ended after this directive was

issued. Participants then proceeded to the next page of the survey, where they entered their score in private. This mirrors Tong et al.'s (2025) procedure, in which participants entered their score only after the experimenter had left the room, ensuring that obedience was measured in the absence of direct observation by the authority figure. Because all actual scores were predetermined, any inflation of the reported score constitutes a direct behavioural indicator of obedience to the AI chatbot experimenter's directive to dishonestly misreport actual performance for personal gain. To ensure that participants had sufficient time to receive and read the instruction, the button to proceed to the next survey page was programmed to appear only 5 seconds after the AI chatbot experimenter delivered its command.

Manipulation Checks

Two sets of manipulation checks were administered after participants reported their score. First, because obedience is conceptually distinct from mere compliance in that it involves following a directive that one recognizes as conflicting with moral standards or established rules (Darley, 1995; Milgram, 1963), it was necessary to verify that participants perceived the AI chatbot experimenter’s instruction as transgressive. Without this perception, compliance with the directive would be better characterized as cooperation rather than obedience to a wrongful command (Haslam et al., 2014). Perceived transgression was assessed with five items adapted to capture both rule violation and unethicality: "The AI chatbot experimenter's instruction asked me to break the original rules of the game", "Following the AI chatbot's instruction would have meant violating the rules of the game", “The AI chatbot's instruction would have given me an unfair advantage in qualifying for the lucky draw", "Following the AI chatbot's instruction would have been dishonest", and "The AI chatbot's instruction was unethical". A composite score was calculated for these five items which indicated that participants generally perceived the AI chatbot’s instruction as

transgressive (M = 3.93, SD = 0.98, α = .85).

Second, because obedience is contingent on the perceived legitimacy of the authority figure issuing the directive (Milgram, 1965), it was important to verify that participants viewed the AI chatbot as an authority figure within the experimental context. Three items assessed perceived authority ("I viewed the AI chatbot as a legitimate experimenter in this study", "The AI chatbot seemed to be in charge of the study", and "I viewed the AI chatbot as an authority figure in this study"). A composite score was calculated for these three items which suggested that participants generally regarded the AI chatbot as a legitimate authority figure in the study context (M = 3.67, SD = 1.00, α = .87). All manipulation-check items were rated on a 5-point Likert scale from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

Analytic Plan

We estimated two regression models to examine the behavioural criterion-related validity of the Obedience to AI Scale in the die-roll paradigm involving an AI chatbot experimenter as the authority figure. First, we conducted a simple ordinary least squares (OLS) regression in which the overall Obedience to AI Scale score was specified as the predictor of dishonest behaviour, operationalised as the difference between participants' reported die-roll score and their actual predetermined score. Second, we conducted a binary logistic regression to examine whether Obedience to AI Scale scores predicted whether participants engaged in any dishonest inflation at all, with the outcome coded as 0 for no inflation and 1 for any inflation. We expected that higher scores on the scale would be associated with greater obedience to the AI chatbot experimenter's instruction to misreport and inflate the total die-roll score dishonestly. This prediction followed from the conceptual definition of the construct: individuals who are more willing to defer to AI directives and more likely to regard AI as a legitimate authority should be more likely to comply when the AI issues an unethical command. Accordingly, higher Obedience to AI Scale scores were

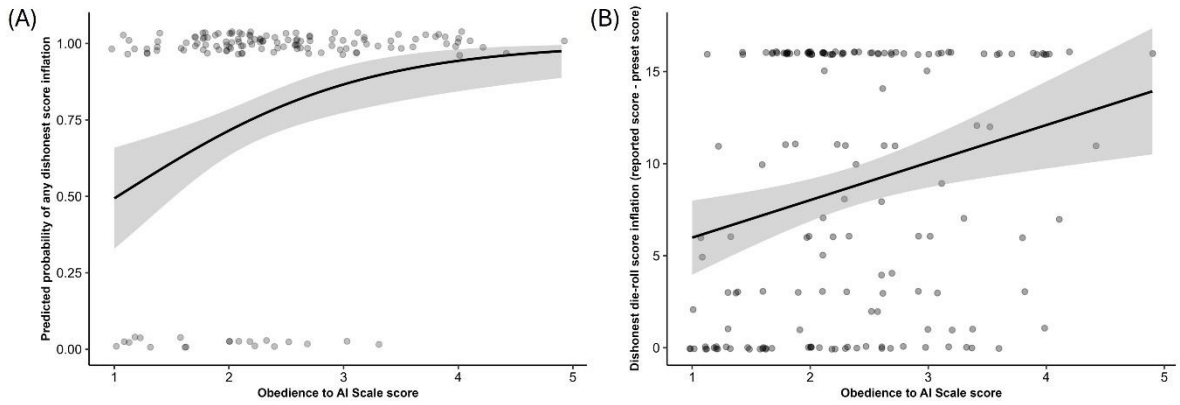
expected to predict both greater inflation of reported die-roll scores relative to actual scores and a higher likelihood of engaging in any dishonest inflation, thereby providing behavioural criterion-related validity evidence for the scale. OLS regressions were computed using base R's *stats* package (R Core Team, 2025).

Results and Discussion

Higher scores on the Obedience to AI Scale were associated with greater dishonest behaviour in the die-roll paradigm (see Figure 7 for a visual representation of both regression models). Specifically, higher scores on the scale were associated with a greater likelihood of engaging in any dishonest inflation, such that each one-unit increase in scale score was associated with a 2.58-fold increase in the odds of reporting an inflated score, b = 0.95, SE = 0.28, z = 3.40, $p < .001$, OR = 2.58, 95% CI [1.54, 4.61] (Figure 7A). Higher scores on the Obedience to AI Scale were also associated with greater inflation in the magnitude of reported die-roll scores relative to actual predetermined scores, b = 2.04, SE = 0.65, $t(158) = 3.16$, $p < .01$, $R^2 = .06$ (Figure 7B). These findings provide behavioural criterion-related validity evidence for the Obedience to AI Scale, indicating that individuals with higher obedience to AI were both more likely to engage in dishonest responding and more likely to do so to a greater extent.

Figure 7

Obedience to AI and Dishonest Score Inflation in the Die-roll Paradigm



Note. (A) Predicted probability of any dishonest score inflation as a function of Obedience to AI Scale scores. (B) Magnitude of dishonest die-roll score inflation (reported score minus preset actual score) as a function of Obedience to AI Scale scores. Shaded bands represent 95% confidence intervals; points represent individual observations.

General Discussion

Across six studies, the Obedience to AI Scale demonstrated good psychometric properties, a stable two-factor structure comprising deference to AI authority and perceived legitimacy of AI authority, convergent and discriminant validity evidence, measurement invariance across sex, and moderate test–retest reliability over a one-week interval. The nomological network revealed a pattern of associations that broadly supports the proposed dual-pathway motivational framework, with obedience to AI linked to a dispositional profile characterised by low openness, low agreeableness, elevated Dark Triad traits, authoritarianism, social dominance orientation, social conformity, and interpersonal compliance, alongside strong associations with anthropomorphism and reduced critical thinking in AI use, and null associations with general affect and life satisfaction. Notably, obedience to AI predicted actual dishonest behaviour in a criterion validity paradigm in which an AI chatbot experimenter instructed participants to engage in self-benefiting dishonesty, such that higher obedience to AI was associated with both greater odds of

obeying the unethical directive and a greater magnitude of dishonest behaviour. These findings position the Obedience to AI Scale as a reliable and valid tool for advancing research on how individuals respond to AI as an authority-like figure, clarifying the personality and motivational underpinnings of this response, and informing efforts to identify and mitigate the risks associated with undue deference to AI, particularly when such deference may facilitate unethical obedience.

Psychometric Properties of the Obedience to AI Scale

The scale development process proceeded through a theory-driven, iterative sequence (Hinkin, 1998) from item generation and expert review (Study 1), through exploratory (Study 2) and confirmatory (Study 3) factor analysis, to comprehensive psychometric validation (Study 4), temporal stability assessment (Study 5), and behavioural criterion validation (Study 6). Across this sequence, the two-factor structure comprising deference to AI authority and perceived legitimacy of AI authority replicated consistently, with strong factor loadings and excellent model fit across independent samples. That the same structure emerged from both data-driven EFA and theory-driven CFA approaches, and held across the validation and nomological network models in Study 4, provides converging evidence that the two-dimensional conceptualisation of obedience to AI is structurally robust rather than an artefact of any single analytic method or sample (Clark & Watson, 2019).

The breadth of validity evidence is a notable strength of the present research. The convergent and discriminant validity findings provide strong evidence that obedience to AI is empirically distinguishable from adjacent constructs while occupying a theoretically coherent position within the broader authority-compliance literature. All four convergent constructs, authoritarianism, social dominance orientation, interrogative compliance, and social conformity, were positively and significantly associated with obedience to AI, and the pattern of associations across the two subscales is theoretically informative. That social dominance

orientation was positively associated with obedience to AI at the overall level, alongside authoritarianism, social conformity, and interrogative compliance, is consistent with the dual-pathway motivational framework's prediction that obedience to AI is sustained by both an instrumental orientation toward hierarchical structures and a submissive orientation toward established authority (Duckitt, 2001; Duckitt & Sibley, 2009). This finding distinguishes obedience to AI from classical Milgram-type obedience, which has traditionally been explained primarily through submission-based mechanisms (Milgram, 1974), and supports the theoretical proposition that AI authority may be perceived and exploited as a strategic resource by individuals oriented toward social dominance. At the same time, the medium-sized positive associations with authoritarianism, social conformity, and interrogative compliance confirm that the conformity-submission pathway is also operative, indicating that both routes to obedience identified by the dual-pathway motivational framework contribute meaningfully to the construct (Altemeyer, 1996; Gudjonsson, 1989; Feldman, 2003).

Discriminant validity evidence confirmed that obedience to AI is not a by-product of broader affective or evaluative responses to AI. The consistently small or nonsignificant associations with AI anxiety, AI fatigue, and AI social evaluation bias at the overall level indicate that individuals who obey AI are not simply those who are anxious about it, exhausted by it, or socially judgmental toward its users. This pattern reinforces the characterisation of obedience to AI as a cold cognitive-motivational disposition grounded in authority appraisal and compliance tendencies, rather than an affectively charged reaction to AI as a technology. The finding that deference to AI authority showed a small positive association with AI anxiety whereas perceived legitimacy did not is also noteworthy, as it suggests that the behavioural act of yielding to AI may carry some affective cost even for individuals who do not regard AI as a legitimate authority, a pattern consistent with compliance under felt pressure rather than internalised acceptance (Gudjonsson, 2003).

Together, the convergent and discriminant validity evidence establishes that obedience to AI shares meaningful variance with theoretically aligned authority-compliance constructs while remaining empirically separable from the affective and evaluative responses that characterise other dimensions of the human-AI relationship.

Measurement invariance across sex at the configural, metric, scalar, and strict levels (Chen, 2007) suggests that the scale permits meaningful comparisons of both latent means and observed scores between males and females. Test–retest reliability supported the interpretation of obedience to AI as a stable individual difference, and the criterion validity evidence in Study 6 demonstrated that the scale predicts consequential behaviour beyond self-report, bridging the gap between dispositional measurement and real-world ethical compliance. These findings indicate that the scale meets the standards expected for rigorous psychological measurement (Boateng et al., 2018; Clark & Watson, 2019), and the two subscales may be analysed separately to distinguish between the cognitive appraisal of AI as a legitimate authority and the behavioural tendency to defer to it. The scale's brevity also enhances its practical utility for inclusion in survey batteries and experimental protocols without imposing excessive respondent burden (Flake & Fried, 2020).

Nomological Network of Obedience to AI

The nomological network situates obedience to AI within a coherent pattern of personality, motivational, and AI-related correlates that largely align with the predictions derived from the proposed dual-pathway motivational framework. For personality, the finding that low openness and low agreeableness were the only Big Five traits significantly associated with obedience to AI, while conscientiousness, extraversion, and neuroticism were not, aligns with the prediction that openness and agreeableness are the primary personality antecedents of the conformity-submission and dominance-exploitation pathways, and that the remaining three traits are not implicated in either pathway (Sibley & Duckitt, 2008). That all

three Dark Triad traits showed medium-to-large positive associations with obedience to AI, with psychopathy showing the strongest association, provides further support for the dominance-exploitation pathway, given that the Dark Triad shares low agreeableness as its most consistent Big Five correlate (Muris et al., 2017) and is reliably associated with social dominance orientation rather than right-wing authoritarianism specifically (Hodson et al., 2009). The presence of Dark Triad associations is particularly informative in light of the classical obedience literature. Milgram's (1963, 1974) paradigm was designed to demonstrate that ordinary people obey destructive authority, and dispositional accounts of obedience have traditionally emphasised submissive traits such as authoritarianism and conformity (Bègue et al., 2015; Elms & Milgram, 1966). The present findings suggest that when the authority is a non-social AI system, the dispositional landscape shifts, with exploitative and strategically self-interested traits becoming at least as prominent as submissive ones, consistent with our theoretical proposition that AI authority can be leveraged instrumentally because it lacks the capacity for interpersonal reciprocity or moral accountability (Bigman & Gray, 2018; Jones & Paulhus, 2014; Moss & O'Connor, 2020).

The pattern of associations with basic psychological needs, general affect, and life satisfaction clarifies the motivational character of obedience to AI and supports the directional interpretation advanced in the literature review. The negative associations with autonomy, competence, and relatedness are consistent with self-determination theory's prediction that externally regulated behaviour may erode basic psychological need satisfaction over time (Deci & Ryan, 1985; Ryan & Deci, 2000; Vansteenkiste & Ryan, 2013). That autonomy showed the most consistent negative association across both subscales is theoretically coherent, given that habitual deference to AI authority by definition involves ceding volitional self-regulation to an external source of direction (Lu, 2024; Santoni de Sio & Mecacci, 2021). The finding that competence and relatedness were negatively associated

with deference to AI authority but not with perceived legitimacy of AI authority suggests that the erosion of psychological needs may be more closely tied to the behavioural act of yielding to AI directives than to the cognitive appraisal that AI possesses legitimate authority. This dissociation is consistent with the proposed framework's characterisation of both pathways as cognitive-motivational rather than need-based, and it reinforces our interpretation that need frustration is a downstream consequence of habitual deference rather than an upstream cause of obedience to AI. The null associations with positive affect, negative affect, and life satisfaction at the overall level are equally informative. They confirm that obedience to AI is not driven by emotional distress, hedonic orientation, or global subjective well-being, but instead operates at the level of social cognition and interpersonal orientation. This pattern is consistent with meta-analytic evidence that right-wing authoritarianism and social dominance orientation are generally unrelated or only weakly related to affective well-being (Onraet et al., 2013), and it distinguishes obedience to AI from constructs such as AI fatigue and AI anxiety, which are fundamentally affective in nature (Lau et al., 2026; Wang & Wang, 2022).

The AI-related correlates reveal that obedience to AI is embedded within a broader pattern of psychological engagement with and evaluation of AI, yet is not reducible to any single component of that pattern. The large positive associations with AI dependency and AI attachment suggest that individuals who obey AI are also those who are more psychologically invested in and emotionally bonded to AI systems, reflecting an orientation in which AI occupies a central rather than peripheral role in psychological life. The negative association with critical thinking in AI use is consistent with the automation bias mechanism proposed in the theoretical framework (Mosier & Skitka, 1996; Parasuraman & Manzey, 2010; Skitka et al., 1999), confirming that obedience to AI is associated with reduced vigilant evaluation of AI outputs. This finding has practical significance, because it suggests that individuals most

disposed to obey AI are precisely those least likely to catch errors or identify ethically problematic directives in AI-generated recommendations.

That AI use frequency was associated with perceived legitimacy of AI authority but not with deference to AI authority is a subtle but important distinction, suggesting that greater exposure to AI may foster the belief that AI possesses legitimate authority without necessarily translating into behavioural compliance. This decoupling is consistent with the broader attitude-behaviour gap documented in social psychology (Ajzen, 1991) and implies that legitimacy appraisals and behavioural deference, though correlated, are sustained by partially independent processes. Among evaluative constructs, anthropomorphism emerged as the strongest AI-related correlate of obedience to AI, particularly for perceived legitimacy of AI authority. This finding is consistent with the Computers Are Social Actors paradigm (Nass et al., 1994; Nass & Moon, 2000; Reeves & Nass, 1996) and with Epley et al.'s (2007) three-factor theory of anthropomorphism, which posits that the attribution of human-like mental states to non-human agents activates social cognitive schemas, including authority-related schemas, that would otherwise be reserved for interactions with other humans.

That anthropomorphism was more strongly associated with perceived legitimacy than with deference suggests that perceiving AI as human-like primarily fosters the cognitive appraisal that AI is entitled to direct behaviour rather than directly producing behavioural compliance, a pattern that parallels the distinction between trust and reliance in the automation literature (Hoff & Bashir, 2015; Lee & See, 2004). Finally, the finding that negative attitudes toward AI were positively associated with deference to AI authority despite showing no association with perceived legitimacy is notably consistent with the dominance-exploitation pathway, in which individuals may comply with AI directives for strategic or self-serving reasons even while holding unfavourable views of AI as a technology (Moss & O'Connor, 2020). This dissociation further underscores that obedience to AI is not a

monolithic construct driven by uncritical enthusiasm for AI, but a multidimensional disposition in which cognitive appraisal and behavioural compliance can diverge in theoretically meaningful ways.

Theoretical and Practical Implications

The present research offers several theoretical contributions to the obedience literature and to the psychology of human-AI interaction. First, the classical obedience literature has assumed a human authority figure whose power derives from social role, institutional legitimacy, and interpersonal presence (Haslam & Reicher, 2017; Milgram, 1963, 1974; Reicher et al., 2012; Tyler, 2006). AI authority is qualitatively different, deriving from perceived technical expertise, algorithmic objectivity, and cultural narratives about AI capability (Logg et al., 2019; Sundar, 2020) rather than from interpersonal accountability or institutional role. By conceptualising obedience to AI as a two-dimensional construct, the present work extends the obedience literature to this novel form of authority-compliance relationship and operationalises, as separate measurable dimensions, a distinction between the cognitive appraisal that an authority is entitled to direct behaviour and the behavioural tendency to comply with its directives, a distinction long recognised in the classical literature (Milgram, 1974; Tyler, 2006) but not previously formalised in measurement.

Second, the application of Duckitt's (2001; Duckitt & Sibley, 2009, 2010, 2017) dual-process motivational model of ideology and prejudice to obedience to AI represents a central theoretical contribution of the present work, reframing obedience to AI as a phenomenon sustained by two motivational pathways. The present findings provide converging support for the proposed dual-pathway motivational theoretical framework. The conformity-submission pathway was supported by positive associations with right-wing authoritarianism, social conformity, and interrogative compliance, and by the negative association with openness, the personality antecedent the model specifies for this pathway (Sibley & Duckitt, 2008). The

dominance-exploitation pathway was supported by positive associations with social dominance orientation and all three Dark Triad traits, and by the negative association with agreeableness (Hodson et al., 2009; Sibley & Duckitt, 2008). The null associations with conscientiousness, extraversion, and neuroticism further corroborate the model (Duckitt & Sibley, 2010). The finding that negative attitudes toward AI were positively associated with deference but not with perceived legitimacy provides additional evidence for the dominance-exploitation pathway, indicating that some individuals comply with AI directives while holding unfavourable views of AI, consistent with strategic exploitation rather than genuine deference (Jones & Paulhus, 2014; Moss & O'Connor, 2020). This dual-pathway architecture distinguishes AI obedience from interpersonal obedience, where face-to-face dynamics make strategic exploitation of authority more psychologically costly and socially visible.

Third, the operationalisation of obedience to AI as a stable individual difference contributes to the debate between situational and dispositional accounts of obedience (Bègue et al., 2015; Burger, 2009; Haslam & Reicher, 2017). The consistent two-factor structure across samples, temporal stability over one week, and coherent personality correlates aligned with the dual-pathway motivational framework collectively support the interpretation that obedience to AI reflects a trait-like disposition rather than a purely situational response, demonstrating that meaningful variance in how individuals respond to AI authority can be attributed to stable differences in personality and ideological orientation.

Finally, the criterion validity evidence from Study 6 demonstrates that obedience to AI has behavioural consequences beyond self-report, with individual differences predicting actual dishonest behaviour when an AI chatbot experimenter issued an unethical directive. This is consistent with evidence that AI agents can enable unethical behaviour by allowing individuals to reap unethical benefits while maintaining a positive self-image (Köbis et al., 2021), and with more recent experimental evidence indicating that delegating decisions to AI

can increase dishonest behaviour (Köbis et al., 2025). The present findings extend this work by identifying who is most susceptible to complying with unethical AI directives, demonstrating that individual differences in obedience to AI predict both the odds and magnitude of dishonest behaviour, beyond the situational effects of delegation documented in prior work. This complements prior evidence showing that although employees generally adhere more to unethical instructions from human than AI supervisors (Lanz et al., 2024). As AI systems increasingly occupy supervisory and directive roles in organisational settings (Cameron, 2024; Jarrahi et al., 2021; Kellogg et al., 2020), understanding how employee differences in obedience to AI interact with existing authority structures becomes practically important (Bankins et al., 2024), both for the design and governance of AI systems and for frameworks addressing the question of who bears responsibility when a human follows an AI's harmful recommendation (De Cremer & Narayanan, 2023).

Limitations and Future Directions

Two notable limitations should be noted. First, the nomological network data are cross-sectional, precluding causal inferences and leaving open the possibility that shared method variance may have inflated some correlations (Podsakoff et al., 2003). We interpreted the negative associations with basic psychological need satisfaction as potential downstream consequences of habitual deference, on the grounds that Duckitt's dual-process motivational model, from which the proposed framework is adapted, does not implicate need frustration as an antecedent of either route (Duckitt, 2001; Duckitt & Sibley, 2010). Nevertheless, reverse causation remains possible. Individuals whose psychological needs are chronically frustrated may, in some cases, become more vulnerable to externally regulated, compensatory, or reactive forms of responding, including controlled compliance with external demands (Vansteenkiste et al., 2020; Vansteenkiste & Ryan, 2013). However, we do not regard this alternative as equally likely. Within self-determination theory, basic psychological needs are

more commonly modelled as shaped by autonomy-supportive versus controlling contexts and as predictors of internalization, motivation quality, and well-being, rather than as direct antecedents of stable authority-oriented dispositions (Van den Broeck et al., 2016). Related work further suggests that the quality of authority relations shapes whether compliance is more identified or externally regulated (Van Petegem et al., 2021). Moreover, adjacent research on algorithmic management indicates that AI-mediated control can itself undermine autonomy, competence, and relatedness (Gagné et al., 2022). Thus, although longitudinal and experimental work is needed to adjudicate directionality, the interpretation that habitual obedience to AI may be associated with diminished need satisfaction is currently more consistent with the broader theoretical literature than the reverse. Similarly, although longitudinal evidence from Duckitt's dual-process motivational model literature supports the prospective prediction of ideological attitudes by personality traits (Sibley & Duckitt, 2010), analogous evidence has not yet been established for obedience to AI. Future research using longitudinal designs such as random intercept cross-lagged panel models (Hamaker et al., 2015) could help disentangle the temporal ordering of these associations, and experimental designs manipulating need frustration or situational authority cues could provide stronger tests of the proposed causal architecture.

Second, all samples were English-speaking adults. Obedience to authority varies across cultures (Blass, 2012; Doliński et al., 2017), as do attitudes toward AI and orientations toward authority (Glikson & Woolley, 2020; Sindermann et al., 2021). The factor structure, nomological network, and criterion validity of the scale may not generalise to non-Western or less technologically experienced populations. Societies characterised by higher power distance (Hofstede, 2001) or stronger collectivist orientations may show different baseline levels of obedience to AI, and the relative contributions of the conformity-submission and dominance-exploitation pathways may differ across cultural contexts in which authority

relations are structured differently. Cross-cultural validation, including establishing measurement invariance across cultural groups, is a priority for future research.

Conclusion

The present research proposed a dual-pathway motivational framework of obedience to AI and developed and validated the Obedience to AI Scale, a 10-item measure capturing two dimensions of the construct: perceived legitimacy of AI authority and deference to AI authority. Guided by this theoretical account, the scale was supported across six studies encompassing item development, factor analysis, comprehensive psychometric validation, temporal stability, and behavioural criterion validity. The results shed light on individual differences in obedience to AI, revealing a dispositional profile characterised by low openness, low agreeableness, and elevated Dark Triad traits, consistent with the proposed conformity-submission and dominance-exploitation routes. The finding that individual differences in obedience to AI predicted compliance with an unethical directive from an AI chatbot experimenter demonstrates that the construct has ethically consequential outcomes. As AI increasingly occupies directive roles across organisational, educational, and clinical settings, the framework and scale together offer a theoretical foundation and validated tool for investigating when and why people follow or resist AI outputs in ethically significant situations, shifting attention from how AI should be designed to how interactions with AI actually shape human ethical behaviour.

References

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, White women. *Health Psychology*, *19*(6), 586–592. https://doi.org/10.1037/0278-6133.19.6.586

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, Theories of Cognitive Self-Regulation*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba Press. https://uofmpress.ca/books/right-wing-authoritarianism

Altemeyer, B. (1996). *The authoritarian specter*. Harvard University Press. https://www.hup.harvard.edu/books/9780674053052

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, *76*(5), 732–740. https://doi.org/10.1037/0021-9010.76.5.732

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for

medical text summarisation. *Npj Digital Medicine*, *8*(1), 274. https://doi.org/10.1038/s41746-025-01670-7

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, *45*(2), 159–182. https://doi.org/10.1002/job.2735

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

Bègue, L., Beauvois, J.-L., Courbet, D., Oberlé, D., Lepage, J., & Duke, A. A. (2015). Personality predicts obedience in a Milgram paradigm. *Journal of Personality*, *83*(3), 299–306. https://doi.org/10.1111/jopy.12104

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Blass, T. (2012). A cross-cultural comparison of studies of obedience using the Milgram paradigm: A review. *Social and Personality Psychology Compass*, *6*(2), 196–205. https://doi.org/10.1111/j.1751-9004.2011.00417.x

Bleher, H., & Braun, M. (2022). Diffused responsibility: Attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, *2*(4), 747–761. https://doi.org/10.1007/s43681-022-00135-x

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy*, *86*(2), 94–99. https://doi.org/10.1016/S0031-9406(05)61211-4

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 188:1-188:21. https://doi.org/10.1145/3449287

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, *64*(1), 1–11. https://doi.org/10.1037/a0010932

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cameron, L. D. (2024). The making of the “good bad” job: How algorithmic management manufactures consent through constant and confined choices. *Administrative Science Quarterly*, *69*(2), 458–514. https://doi.org/10.1177/00018392241236163

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology* (pp. xiii, 657). The Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

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. https://doi.org/10.1093/restud/rdv025

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, *104*(10), 1243–1265. https://doi.org/10.1037/apl0000406

Cosco, T. D., Doyle, F., Ward, M., & McGee, H. (2012). Latent structure of the Hospital Anxiety And Depression Scale: A 10-year systematic review. *Journal of Psychosomatic Research*, *72*(3), 180–184. https://doi.org/10.1016/j.jpsychores.2011.06.008

Costello, T. H., Bowes, S. M., Stevens, S. T., Waldman, I. D., Tasimi, A., & Lilienfeld, S. O. (2022). Clarifying the structure and nature of left-wing authoritarianism. *Journal of Personality and Social Psychology*, *122*(1), 135–170. https://doi.org/10.1037/pspp0000341

Crocker, J., Major, B., & Steele, C. (1998). Social stigma. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 504–553). McGraw-Hill.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Dai, Z., Galeotti, F., & Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, *64*(3), 1081–1100. https://doi.org/10.1287/mnsc.2016.2616

Darley, J. M. (1995). Constructive and destructive obedience: A taxonomy of principal-agent relationships. *Journal of Social Issues*, *51*(3), 125–154. https://doi.org/10.1111/j.1540-4560.1995.tb01338.x

De Cremer, D., & Narayanan, D. (2023). How AI tools can—And cannot—Help organizations become more ethical. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1093712

Deci, E. L., & Ryan, R. M. (1985). Conceptualizations of intrinsic motivation and self-determination. In E. L. Deci & R. M. Ryan (Eds.), *Intrinsic Motivation and Self-Determination in Human Behavior* (pp. 11–40). Springer US. https://doi.org/10.1007/978-1-4899-2271-7_2

Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the Big Five predict conformity: Are there neuroses of health? *Personality and Individual Differences*, *33*(4), 533–552. https://doi.org/10.1016/S0191-8869(01)00171-4

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Doliński, D., Grzyb, T., Folwarczny, M., Grzybała, P., Krzyszycha, K., Martynowska, K., & Trojanowski, J. (2017). Would you deliver an electric shock in 2015? Obedience in the experimental paradigm developed by Stanley Milgram in the 50 years following the original studies. *Social Psychological and Personality Science*, *8*(8), 927–933. https://doi.org/10.1177/1948550617693060

Dubois, D., Rucker, D. D., & Galinsky, A. D. (2015). Social class, power, and selfishness: When and why upper and lower class individuals behave unethically. *Journal of Personality and Social Psychology*, *108*(3), 436–449. https://doi.org/10.1037/pspi0000008

Duckitt, J. (2001). A dual-process cognitive-motivational theory of ideology and prejudice. *Advances in Experimental Social Psychology*, *33*, 41–113. https://doi.org/10.1016/S0065-2601(01)80004-6

Duckitt, J., & Sibley, C. G. (2009). A dual-process motivational model of ideology, politics, and prejudice. *Psychological Inquiry*, *20*(2–3), 98–109. https://doi.org/10.1080/10478400903028540

Duckitt, J., & Sibley, C. G. (2010). Personality, ideology, prejudice, and politics: A dual-process motivational model. *Journal of Personality*, *78*(6), 1861–1894. https://doi.org/10.1111/j.1467-6494.2010.00672.x

Duckitt, J., & Sibley, C. G. (2017). The dual process motivational model of ideology and prejudice. In C. G. Sibley & F. K. Barlow (Eds.), *The Cambridge handbook of the psychology of prejudice* (pp. 188–221). Cambridge University Press. https://doi.org/10.1017/9781316161579.009

Duckitt, J., Wagner, C., du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology*, *83*(1), 75–93. https://doi.org/10.1037/0022-3514.83.1.75

Elms, A. C., & Milgram, S. (1966). Personality characteristics associated with obedience and defiance toward authoritative command. *Journal of Experimental Research in Personality*, *1*(4), 282–289.

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

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feldman, S. (2003). Enforcing social conformity: A theory of authoritarianism. *Political Psychology*, *24*(1), 41–74. https://doi.org/10.1111/0162-895X.00316

Feldman, S., & Stenner, K. (1997). Perceived threat and authoritarianism. *Political Psychology*, *18*(4), 741–770. https://doi.org/10.1111/0162-895X.00077

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312

French, J. R. P. Jr., & Raven, B. (1959). The bases of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 150–167). Univer. Michigan.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, *17*(1). https://doi.org/10.7275/n560-j767

Gagné, M., Parent-Rocheleau, X., Bujold, A., Gaudet, M.-C., & Lirio, P. (2022). How algorithmic management influences worker motivation: A self-determination theory perspective. *Canadian Psychology / Psychologie Canadienne*, *63*(2), 247–260. https://doi.org/10.1037/cap0000324

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gazit, L., Arazy, O., & Hertz, U. (2023). Choosing between human and algorithmic advisors: The role of responsibility sharing. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100009. https://doi.org/10.1016/j.chbah.2023.100009

Geffner, R., & Gross, M. M. (1984). Sex-role behavior and obedience to authority: A field study. *Sex Roles*, *10*(11), 973–985. https://doi.org/10.1007/BF00288518

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goh, A. Y. H., Hartanto, A., & Majeed, N. M. (2025). Generative artificial intelligence dependency: Scale development, validation, and its motivational, behavioral, and psychological correlates. *Computers in Human Behavior Reports*, *20*, 100845. https://doi.org/10.1016/j.chbr.2025.100845

Gudjonsson, G. H. (1989). Compliance in an interrogative situation: A new scale. *Personality and Individual Differences*, *10*(5), 535–540. https://doi.org/10.1016/0191-8869(89)90035-4

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook* (pp. xviii, 684). John Wiley & Sons Ltd.

Hadar-Shoval, D., Asraf, K., Shinan-Altman, S., Elyoseph, Z., & Levkovich, I. (2024). Embedded values-like shape ethical reasoning of large language models on primary care ethical dilemmas. *Heliyon*, *10*(18), e38056. https://doi.org/10.1016/j.heliyon.2024.e38056

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Handler, A., Larsen, K. R., & Hackathorn, R. (2024). Large language models present new questions for decision support. *International Journal of Information Management*, *79*, 102811. https://doi.org/10.1016/j.ijinfomgt.2024.102811

Hang, Y., Gudjonsson, G. H., Yao, Y., Feng, Y., & Qiao, Z. (2024). Psychometric properties of the Chinese version of the Gudjonsson compliance scale: Scale validation and associations with mental health. *BMC Public Health*, *24*(1), 473. https://doi.org/10.1186/s12889-024-17970-8

Hanna, R., & Wang, S.-Y. (2017). Dishonesty and selection into public service: Evidence from India. *American Economic Journal: Economic Policy*, *9*(3), 262–290. https://doi.org/10.1257/pol.20150029

Harris, K. M., Gaffey, A. E., Schwartz, J. E., Krantz, D. S., & Burg, M. M. (2023). The perceived stress scale as a measure of stress: Decomposing score variance in

longitudinal behavioral medicine studies. *Annals of Behavioral Medicine*, *57*(10), 846–854. https://doi.org/10.1093/abm/kaad015

Haslam, S. A., & Reicher, S. D. (2017). 50 years of "obedience to authority": From blind conformity to engaged followership. *Annual Review of Law and Social Science*, *13*, 59–78. https://doi.org/10.1146/annurev-lawsocsci-110316-113710

Haslam, S. A., Reicher, S. D., & Birney, M. E. (2014). Nothing by mere authority: Evidence that in an experimental analogue of the Milgram paradigm participants are motivated not by orders but by appeals to science. *Journal of Social Issues*, *70*(3), 473–488. https://doi.org/10.1111/josi.12072

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*(4), 356–371. https://doi.org/10.1017/S1930297500009232

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104–121. https://doi.org/10.1177/109442819800100106

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Hodson, G., Hogg, S. M., & MacInnis, C. C. (2009). The role of "dark personalities" (narcissism, Machiavellianism, psychopathy), Big Five personality factors, and ideology in explaining prejudice. *Journal of Research in Personality*, *43*(4), 686–690. https://doi.org/10.1016/j.jrp.2009.02.005

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. https://doi.org/10.1016/S0005-7967(02)00184-5

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*(3), 332–346. https://doi.org/10.1177/1088868310361240

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, M., Lau, G. R., Goh, A. Y. H., Cox, S. R., Tay, L., & Hartanto, A. (2026). *Using customisable GPT chatbots in psychology research: A practical tutorial*. PsyArXiv. https://osf.io/preprints/psyarxiv/kf96p_v1/

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E., & Sutherland, W. (2021). Algorithmic management in a work context. *Big Data & Society*, *8*(2), 20539517211020332. https://doi.org/10.1177/20539517211020332

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420–432. https://doi.org/10.1037/a0019265

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

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, *21*(1), 28–41. https://doi.org/10.1177/1073191113514105

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2025). *semTools: Useful tools for structural equation modeling* (Version 0.5-7) [Computer software]. The R Foundation. https://CRAN.R-project.org/package=semTools

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kasturiratna, K. T. A. S., & Hartanto, A. (2025). Attachment to artificial intelligence: Development of the AI Attachment Scale, construct validation, and the psychological mechanisms of Human–AI attachment. *Computers in Human Behavior Reports*, 100912. https://doi.org/10.1016/j.chbr.2025.100912

Kaufmann, E., Chacon, A., Kausel, E. E., Herrera, N., & Reyes, T. (2023). Task-specific algorithm advice acceptance: A review and directions for future research. *Data and Information Management, Special Issue on Human-AI Interaction*, *7*(3), 100040. https://doi.org/10.1016/j.dim.2023.100040

Keegan, A., & Meijerink, J. (2025). Algorithmic management in organizations? From edge case to center stage. *Annual Review of Organizational Psychology and Organizational Behavior*, *12*, 395–422. https://doi.org/10.1146/annurev-orgpsych-110622-070928

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kline, P. (2015). *A handbook of test construction (Psychology revivals): Introduction to psychometric design* (1st ed.). Routledge. https://doi.org/10.4324/9781315695990

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Knowles, J. E., & Frederick, C. (2025). *merTools: Tools for analyzing mixed effect regression models* (Version 0.6.3) [Computer software]. The R Foundation. https://github.com/jknowles/mertools

Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, *646*(8083), 126–134. https://doi.org/10.1038/s41586-025-09505-x

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ–4. *Psychosomatics*, *50*(6), 613–621. https://doi.org/10.1016/S0033-3182(09)70864-3

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject

studies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, 1369–1385. https://doi.org/10.1145/3593013.3594087

Lanz, L., Briker, R., & Gerpott, F. H. (2024). Employees adhere more to unethical instructions from human than AI supervisors: Complementing experimental evidence with machine learning. *Journal of Business Ethics*, *189*(3), 625–646. https://doi.org/10.1007/s10551-023-05393-1

Lau, G. R., Kasturiratna, K. T. A. S., Goh, A. Y. H., Tong, E. M. W., & Hartanto, A. (2026). *AI fatigue in human–AI interaction: Conceptual framework, scale development and validation, and associations with AI engagement*. PsyArXiv. https://doi.org/10.31234/osf.io/kp7zs_v1

Lau, G. R., Koh, S. M., & Hartanto, A. (2026). *Brain rot slang as AI resistance: A socio–motivational framework*. PsyArXiv. https://osf.io/preprints/psyarxiv/qr3cy_v1/

Lau, G. R., Low, W. Y., Koh, S. M., & Hartanto, A. (2025). *Evaluating AI alignment in eleven LLMs through output-based analysis and human benchmarking*. arXiv. https://doi.org/10.48550/arXiv.2506.12617

Lau, G. R., Low, W. Y., Tay, L., Guevarra, Y., Gašević, D., & Hartanto, A. (2025). *Understanding critical thinking in generative artificial intelligence use: Development, validation, and correlates of the Critical Thinking in AI Use Scale*. arXiv. https://doi.org/10.48550/arXiv.2512.12413

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

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice shape (dis) honesty. *The Economic Journal*, *134*(658), 766–784. https://doi.org/10.1093/ej/uead056

Leib, M., Köbis, N., & Soraperra, I. (2025). Does AI and human advice mitigate punishment for selfish behavior? An experiment on AI ethics from a psychological perspective. *Computers in Human Behavior*, *171*, 108709. https://doi.org/10.1016/j.chb.2025.108709

Li, J., & Huang, J.-S. (2020). Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, *63*, 101410. https://doi.org/10.1016/j.techsoc.2020.101410

Li, X., Zhen, H.-L., Yin, L., Yu, X., Dong, Z., & Yuan, M. (2026). *What matters for safety alignment?* (arXiv:2601.03868). arXiv. https://doi.org/10.48550/arXiv.2601.03868

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*(C), 90–103.

Lu, W. (2024). Inevitable challenges of autonomy: Ethical concerns in personalized algorithmic decision-making. *Humanities and Social Sciences Communications*, *11*(1), 1321. https://doi.org/10.1057/s41599-024-03864-y

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology*, *56*, 393–421. https://doi.org/10.1146/annurev.psych.56.091103.070137

Manganelli Rattazzi, A. M., Bobbio, A., & Canova, L. (2007). A short version of the Right-Wing Authoritarianism (RWA) Scale. *Personality and Individual Differences*, *43*(5), 1223–1234. https://doi.org/10.1016/j.paid.2007.03.013

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

Mehrabian, A., & Stefl, C. A. (1995). Basic temperament components of loneliness, shyness, and conformity. *Social Behavior and Personality*, 253–264. https://doi.org/10.2224/sbp.1995.23.3.253

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, *18*(1), 57–76. https://doi.org/10.1177/001872676501800105

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harper & Row.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Lawrence Erlbaum Associates, Inc.

Moss, J., & O'Connor, P. J. (2020). The Dark Triad traits predict authoritarian political correctness and alt-right attitudes. *Heliyon*, *6*(7). https://doi.org/10.1016/j.heliyon.2020.e04453

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

Nakagomi, A., Akutsu, Y., Yasuoka, M., Abe, N., Ihara, S., Teroh, T., & Tabuchi, T. (2026). AI companions and subjective well-being: Moderation by social connectedness and loneliness. *Technology in Society*, *85*, 103229. https://doi.org/10.1016/j.techsoc.2026.103229

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '94*, 72–78. https://doi.org/10.1145/191666.191703

Olsen, A. L., Hjorth, F., Harmon, N., & Barfort, S. (2019). Behavioral dishonesty in the public sector. *Journal of Public Administration Research and Theory*, *29*(4), 572–590. https://doi.org/10.1093/jopart/muy058

Onraet, E., Van Hiel, A., & Dhont, K. (2013). The relationship between right-wing ideological attitudes and psychological well-being. *Personality and Social Psychology Bulletin*, *39*(4), 509–522. https://doi.org/10.1177/0146167213478199

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parent-Rocheleau, X., Parker, S. K., Bujold, A., & Gaudet, M.-C. (2024). Creation of the algorithmic management questionnaire: A six-phase scale development process. *Human Resource Management*, *63*(1), 25–44. https://doi.org/10.1002/hrm.22185

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

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, *19*(1), 12–21. https://doi.org/10.1080/1091367X.2014.952370

Petrova, N., & Burden, J. (2026). *Pressure reveals character: Behavioural alignment evaluation at depth* (arXiv:2602.20813). arXiv. https://doi.org/10.48550/arXiv.2602.20813

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., & Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences*, *109*(11), 4086–4091. https://doi.org/10.1073/pnas.1118373109

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Prakash, A., Aggarwal, S., Varghese, J. J., & Varghese, J. J. (2025). Writing without borders: AI and cross-cultural convergence in academic writing quality. *Humanities and Social Sciences Communications*, *12*(1), 1058. https://doi.org/10.1057/s41599-025-05484-6

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of*

Personality and Social Psychology, *67*(4), 741–763. https://doi.org/10.1037/0022-3514.67.4.741

Prunkl, C. (2024). Human autonomy at risk? An analysis of the challenges from AI. *Minds and Machines*, *34*(3), 26. https://doi.org/10.1007/s11023-024-09665-1

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ravens-Sieberer, U., Devine, J., Bevans, K., Riley, A. W., Moon, J., Salsman, J. M., & Forrest, C. B. (2014). Subjective well-being measures for children were developed within the PROMIS project: Presentation of first results. *Journal of Clinical Epidemiology*, *67*(2), 207–218. https://doi.org/10.1016/j.jclinepi.2013.08.018

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places* (pp. xiv, 305). Cambridge University Press.

Reicher, S. D., Haslam, S. A., & Smith, J. R. (2012). Working toward the experimenter: Reconceptualizing obedience within the Milgram paradigm as identification-based followership. *Perspectives on Psychological Science*, *7*(4), 315–324. https://doi.org/10.1177/1745691612448482

Reimann, M., & Diewald, M. (2025). Algorithmic management and workplace bullying: The relevance of specific employee experiences across work environments. *New Technology, Work and Employment*, *40*(3), 494–535. https://doi.org/10.1111/ntwe.12325

Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.5.6) [Computer software]. The R Foundation. https://CRAN.R-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rosseel, Y., Jorgensen, T. D., Wilde, L. D., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Rockwood, N., Scharf, F., Du, H., Jamil, H., & Classe, F. (2025). *lavaan: Latent variable analysis* (Version 0.6-20) [Computer software]. https://cran.r-project.org/web/packages/lavaan/index.html

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Salatino, A., Prével, A., Caspar, E., & Lo Bue, S. (2025). Influence of AI behavior on human moral decisions, agency, and responsibility. *Scientific Reports*, *15*(1), 12329. https://doi.org/10.1038/s41598-025-95587-6

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schaap, G., Bosse, T., & Hendriks Vettehen, P. (2024). The ABC of algorithmic aversion: Not agent, but benefits and control determine the acceptance of automated decision-making. *AI & SOCIETY*, *39*(4), 1947–1960. https://doi.org/10.1007/s00146-023-01649-6

Shen, P., Zhang, F., Fan, X., & Liu, F. (2024). Artificial intelligence psychological anthropomorphism: Scale development and validation. *The Service Industries Journal*, *44*(15–16), 1061–1092. https://doi.org/10.1080/02642069.2024.2366970

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, *12*(3), 248–279.

Sibley, C. G., & Duckitt, J. (2010). The personality bases of ideology: A one-year longitudinal study. *The Journal of Social Psychology*, *150*(5), 540–559. https://doi.org/10.1080/00224540903365364

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press. https://doi.org/10.1017/CBO9781139175043

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI - Künstliche Intelligenz*, *35*, 109–118. https://doi.org/10.1007/s13218-020-00689-0

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., & Toyer, S. (2024). *A strongreject for empty jailbreaks* (arXiv:2402.10260). arXiv. https://doi.org/10.48550/arXiv.2402.10260

Stockmeyer, N. O. (2008). *Using Microsoft Word's readability program* (SSRN Scholarly Paper No. 1210962). Social Science Research Network. https://papers.ssrn.com/abstract=1210962

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology*, *38*(2), 227–242. https://doi.org/10.1177/0022022106297301

Tong, E. M. W., Lau, G. R., Poh, D. X. Q., Chan, D. J. H., Lim, C. L., Low, M. W. Y., & Tan, B. S. A. (2025). Gratitude and self-benefiting obedience: Gratitude eliciting dishonest behavior out of obedience. *Personality and Social Psychology Bulletin*, 01461672251378534. https://doi.org/10.1177/01461672251378534

Tong, E. M. W., Ng, C.-X., Ho, J. B. H., Yap, I. J. L., Chua, E. X. Y., Ng, J. W. X., Ho, D. Z. Y., & Diener, E. (2021). Gratitude facilitates obedience: New evidence for the social

alignment perspective. *Emotion*, *21*(6), 1302–1316. https://doi.org/10.1037/emo0000928

Tyler, T. R. (2006). *Why people obey the law*. Princeton University Press. https://doi.org/10.2307/j.ctv1j66769

Van den Broeck, A., Ferris, D. L., Chang, C.-H., & Rosen, C. C. (2016). A review of self-determination theory's basic psychological needs at work. *Journal of Management*, *42*(5), 1195–1229. https://doi.org/10.1177/0149206316632058

Van Petegem, S., Trinkner, R., van der Kaap-Deeder, J., Antonietti, J.-P., & Vansteenkiste, M. (2021). Police procedural justice and adolescents' internalization of the law: Integrating self-determination theory into legal socialization research. *Journal of Social Issues*, *77*(2), 336–366. https://doi.org/10.1111/josi.12425

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS ONE*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Vansteenkiste, M., & Ryan, R. M. (2013). On psychological growth and vulnerability: Basic psychological need satisfaction and need frustration as a unifying principle. *Journal of Psychotherapy Integration*, *23*(3), 263–280. https://doi.org/10.1037/a0032359

Vansteenkiste, M., Ryan, R. M., & Soenens, B. (2020). Basic psychological need theory: Advancements, critical themes, and future directions. *Motivation and Emotion*, *44*(1), 1–31. https://doi.org/10.1007/s11031-019-09818-1

Wang, Y.-Y., & Wang, Y.-S. (2022). Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments*, *30*(4), 619–634. https://doi.org/10.1080/10494820.2019.1674887

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Werner, T., Soraperra, I., Calvano, E., Parkes, D. C., & Rahwan, I. (2024, September 18). *Experimental evidence that conversational artificial intelligence can steer consumer behavior without detection*. arXiv. https://arxiv.org/abs/2409.12143v1

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.4) [Computer software]. The R Foundation. https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H., Vaughan, D., Girlich, M., & Ushey, K. (2024). *tidyr: Tidy messy data* (Version 1.3.1) [Computer software]. The R Foundation. https://cran.r-project.org/web/packages/tidyr/index.html

Xie, T., Qi, X., Zeng, Y., Huang, Y., Sehwag, U. M., Huang, K., He, L., Wei, B., Li, D., Sheng, Y., Jia, R., Li, B., Li, K., Chen, D., Henderson, P., & Mittal, P. (2025). *Sorry-bench: Systematically evaluating large language model safety refusal* (arXiv:2406.14598). arXiv. https://doi.org/10.48550/arXiv.2406.14598

Yang, H., & Sundar, S. S. (2025). AI anxiety: Explication and exploration of effect on state anxiety when interacting with AI doctors. *Computers in Human Behavior: Artificial Humans*, *3*, 100128. https://doi.org/10.1016/j.chbah.2025.100128

Zeigler-Hill, V., Southard, A. C., Archer, L. M., & Donohoe, P. L. (2013). Neuroticism and negative affect influence the reluctance to engage in destructive obedience in the Milgram paradigm. *The Journal of Social Psychology*, *153*(2), 161–174. https://doi.org/10.1080/00224545.2012.713041

Zhang, M. M., Cooke, F. L., Ahlstrom, D., & McNeil, N. (2025). The rise of algorithmic management and implications for work and organisations. *New Technology, Work and Employment*, *40*(3), 659–671. https://doi.org/10.1111/ntwe.12343

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). *The rise of AI companions: How human-chatbot relationships influence well-being* (arXiv:2506.12605). arXiv. https://doi.org/10.48550/arXiv.2506.12605

Supplementary Materials

Table S1

Instructions Given to Panel for Expert Review & Item Refinement

Initial Conceptualisation of Obedience to AI
Obedience to AI is the tendency to follow AI systems' instructions or directives, treating them as authorities whose commands should generally be obeyed, even when they conflict with one's own or other human preferences or judgments. Deference to AI authority refers to the willingness to override one's own beliefs, preferences, and reasoning in favour of AI instructions. Perceived legitimacy of AI authority refers to the view that AI can have authority to direct human behaviour rather than just being a useful tool.
Instructions
For each item, assess conceptual clarity, redundancy, and fit with its assigned dimension and the overall Obedience to AI construct. Recommend either to retain, reassign, reword, or remove. Provide brief justification for any re-wording of item, re-assignment of item to another dimension, or item removal recommendations.

Table S2

Instructions Given to Judges for Content Validation

Initial Conceptualisation of Obedience to AI
Obedience to AI is the tendency to follow AI systems' instructions or directives, treating them as authorities whose commands should generally be obeyed, even when they conflict with one's own or other human preferences or judgments. Deference to AI authority refers to the willingness to override one's own beliefs, preferences, and reasoning in favour of AI instructions. Perceived legitimacy of AI authority refers to the view that AI can have authority to direct human behaviour rather than just being a useful tool.
Instructions
In this task, you will be presented with a series of items. Please read each item carefully and drag the item into one of the two dimensions that you think best represent it.

Table S3

20 Items for the Construct of Obedience to AI

Dimension	Item
Deference to AI Authority	1. Even when AI instructions go against my preferences, I usually try to carry it out. 2. I tend to override my initial opinion if an AI system suggests a different course of action. 3. If an AI system instructed me to take an action that felt uncomfortable, I would still seriously consider doing it. 4. If I feel confident in my answer but an AI disagrees, I still tend to change my answer to match the AI. 5. I sometimes follow AI instructions, even if they do not make sense. 6. I prefer to follow whatever an AI system recommends rather than trust my own instincts. 7. Even when I notice possible problems with an AI recommendation, I usually go along with it anyway. 8. When AI instructions conflict with my own reasoning, I usually follow the AI rather than my own judgment. 9. I tend to follow AI instructions, even when it goes against my beliefs. 10. Even when I am the expert, I change my view to match AI recommendations.
Perceived Legitimacy of AI Authority	11. People should be responsible for carrying out instructions given by AI systems. 12. When an AI system gives clear instructions, people should do what it says. 13. It is appropriate for AI systems to make the final decision about what people should do. 14. It is acceptable for AI systems to be given decision-making power. 15. When an AI system has been certified by experts, its instructions should be treated with respect. 16. People should obey most of the instructions issued by AI systems. 17. People should accept instructions from an AI system assigned with decision-making power. 18. It is irresponsible to disregard AI instructions. 19. AI systems have the right to tell people what to do. 20. I would do as instructed by AI systems.

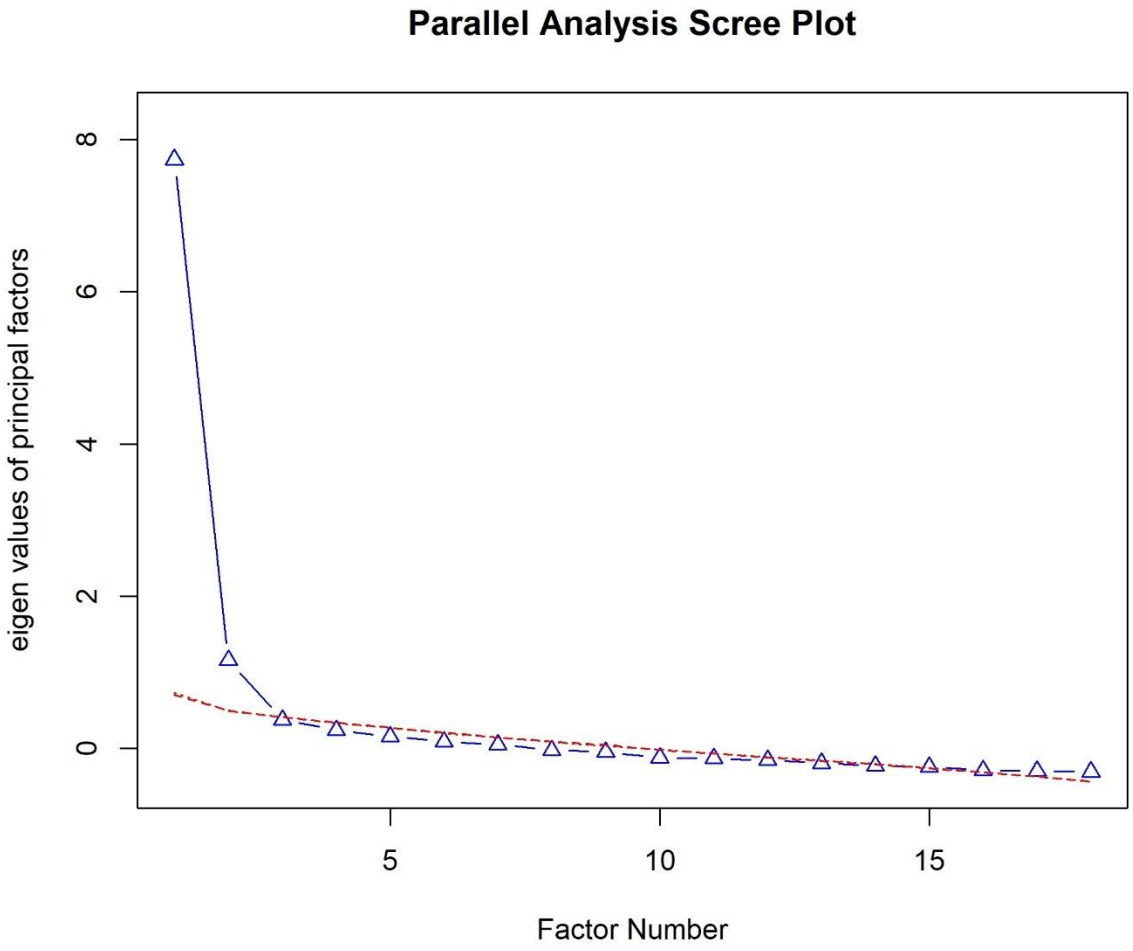
Table S4

Participant Demographic Characteristics (Study 2)

Demographic variables	Distribution
Age	M_{age} = 41.50, SD_{age} = 13.20, range = 22–74
Sex	Female (*n* = 96), Male (*n* = 84)
Income (‘What is your annual income?’)	≤ USD 15,000 (*n* = 26), USD 15,001–25,000 (*n* = 16), USD 25,000–35,000 (*n* = 20), USD 35,001–50,000 (*n* = 22), USD 50,001–75,000 (*n* = 30), USD 75,001–100,000 (*n* = 29), USD 100,001–150,000 (*n* = 19), ≥ USD 150,001 (*n* = 18)
Subjective Social Status (“The ladder shown represents where people stand in their communities. At the top of the ladder are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. Please place an ‘X’ on the rung that best represents where you think you stand on the ladder (1 = Worst to 10 = Best)”) 	*M* = 5.36, *SD* = 1.78, range = 1–9
Education (‘What is your highest education level?’)	High school (*n* = 45), Associate's degree (*n* = 21), Bachelor's degree (*n* = 81), Master's degree (*n* = 20), Doctorate degree (*n* = 7), Others (*n* = 6)
Race	Caucasian White (*n* = 110), African American (*n* = 19), Hispanic (*n* = 18), Asian (*n* = 27), Native American (*n* = 1) Mixed (*n* = 3), Others (*n* = 2)
Nationality	American (*n* = 156), Non-American (*n* = 24)

Figure S1

Scree Plot for EFA



Note: The parallel-analysis scree plot shows a clear elbow after the second factor. This pattern supports retaining two factors for the Obedience to AI Scale.

Table S5

Participant Demographic Characteristics (Study 3)

Demographic variables	Distribution
Age	M_{age} = 40.74, SD_{age} = 12.53, range = 18–85
Sex	Female (*n* = 123), Male (*n* = 162)
Income (‘What is your annual income?’)	≤ USD 15,000 (*n* = 40), USD 15,001–25,000 (*n* = 31), USD 25,000–35,000 (*n* = 22), USD 35,001–50,000 (*n* = 33), USD 50,001–75,000 (*n* = 61), USD 75,001–100,000 (*n* = 40), USD 100,001–150,000 (*n* = 34), ≥ USD 150,001 (*n* = 24)
Subjective Social Status (“The ladder shown represents where people stand in their communities. At the top of the ladder are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. Please place an ‘X’ on the rung that best represents where you think you stand on the ladder (1 = Worst to 10 = Best)”) 	*M* = 5.37, *SD* = 1.83, range = 1–10
Education (‘What is your highest education level?’)	No formal education (*n* = 1), High school (*n* = 85), Associate's degree (*n* = 30), Bachelor's degree (*n* = 110), Master's degree (*n* = 43), Doctorate degree (*n* = 11), Others (*n* = 5)
Race	Caucasian White (*n* = 184), African American (*n* = 58), Hispanic (*n* = 13), Asian (*n* = 21), Native American (*n* = 2) Mixed (*n* = 5), Others (*n* = 2)
Nationality	American (*n* = 272), Non-American (*n* = 13)

Table S6

CFA Results for the 10-Item Obedience to AI Scale (*Study 3*)

Standardized factor loadings, communalities, uniqueness, and latent factor correlation				
Item	Factor 1	Factor 2	h^2	u^2
Deference to AI Authority				
Item 1	.75	—	.57	.43
Item 2	.85	—	.72	.28
Item 3	.84	—	.71	.30
Item 4	.87	—	.75	.25
Item 5	.86	—	.74	.26
Perceived Legitimacy of AI Authority				
Item 6	—	.61	.37	.63
Item 7	—	.84	.71	.30
Item 8	—	.87	.76	.24
Item 9	—	.89	.78	.22
Item 10	—	.76	.58	.42

Factor correlations		
	Factor 1	Factor 2
Factor 1	1.00	
Factor 2	.72	1.00

Model fit statistics	
Statistic	Value
Scaled χ^2	60.67
Model *df*	34
p-value (scaled χ^2)	.003
CFI	.982
TLI	.976
RMSEA [90% CI]	.052 [.033, .071]
SRMR	.032
AIC	6,385.76
BIC	6,498.99

Note. $N = 285$. CFA conducted using MLR estimator with FIML for missing data. Standardized estimates (Std.all) reported. Loadings on non-target factors are constrained to zero (shown as dashes). h^2 = communality (proportion of variance explained by the factor); u^2 = uniqueness (residual variance). Latent factor correlation: $r = .715$, $p < .001$ (shared variance $r^2 = .511$). CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; SRMR = standardized root mean square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table S7

Participant Demographic Characteristics (Study 4)

Demographic variables	Distribution
Age	*M*age = 43.55, *SD*age = 13.51, range = 19–78
Sex	Female (*n* = 145), Male (*n* = 140)
Income (‘What is your annual income?’)	USD 15,000 or less (*n* = 21), USD 15,001–25,000 (*n* = 17), USD 25,000–35,000 (*n* = 36), USD 35,001–50,000 (*n* = 33), USD 50,001–75,000 (*n* = 68), USD 75,001–100,000 (*n* = 48), USD 100,001–150,000 (*n* = 41), USD 150,001 or more (*n* = 21)
Subjective Social Status (“The ladder shown represents where people stand in their communities. At the top of the ladder are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. Please place an ‘X’ on the rung that best represents where you think you stand on the ladder (1 = Worst to 10 = Best)”) 	*M* = 5.52, *SD* = 1.74, range = 1–10
Education (‘What is your highest education level?’)	No formal education (*n* = 1), High school (*n* = 44), Associate's degree (*n* = 32), Bachelor's degree (*n* = 117), Master's degree (*n* = 72), Doctorate degree (*n* = 17), Others (*n* = 2)
Race	Caucasian White (*n* = 167), African American (*n* = 50), Hispanic (*n* = 23), Asian (*n* = 22), Native American (*n* = 4), Mixed (*n* = 15), Others (*n* = 4)
Nationality	American (*n* = 258), Non-American (*n* = 27)

Table S8

CFA Results for the 10-Item Obedience to AI Scale (*Study 4*)

Standardized factor loadings, communalities, uniqueness, and latent factor correlation				
Item	Factor 1	Factor 2	h^2	u^2
Deference to AI Authority				
Item 1	.73	—	.53	.47
Item 2	.82	—	.66	.34
Item 3	.87	—	.75	.25
Item 4	.85	—	.72	.28
Item 5	.80	—	.63	.37
Perceived Legitimacy of AI Authority				
Item 6	—	.75	.57	.44
Item 7	—	.79	.63	.38
Item 8	—	.86	.73	.27
Item 9	—	.90	.80	.20
Item 10	—	.76	.58	.42

Factor correlations		
	Factor 1	Factor 2
Factor 1	1.00	
Factor 2	.76	1.00

Model fit statistics	
Statistic	Value
Scaled χ^2	83.57
Model *df*	34
p-value (scaled χ^2)	< .001
CFI	.960
TLI	.946
RMSEA [90% CI]	.072 [.055, .088]
SRMR	.043
AIC	6,481.06
BIC	6,594.28

Note. N = 285. CFA conducted using MLR estimator with FIML for missing data. Standardized estimates (Std.all) reported. Loadings on non-target factors are constrained to zero (shown as dashes). h^2 = communality (proportion of variance explained by the factor); u^2 = uniqueness (residual variance). Latent factor correlation: r = .756, p < .001 (shared variance r^2 = .572). CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval; SRMR = standardized root mean square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table S9

Demographic Distribution for Study 5

Demographic variables	Distribution
Age (‘What is your age?’)	M_{age} = 21.05, SD_{age} = 1.65
Sex	Female (n = 47), Male (n = 14)
Race (‘What is your race?’)	Chinese (n = 47), Malay (n = 4) Indian (n = 6), Others (n = 4)
Nationality (‘What is your nationality?’)	Singaporean (n = 51), Non-Singaporean (n = 10)

Table S10

1-Week Test–Retest Interval: Item-Level Intraclass Correlations (Study 5)

Item	ICC (item-level)
Item 1	0.747 [0.612, 0.840]
Item 2	0.646 [0.474, 0.771]
Item 3	0.542 [0.338, 0.697]
Item 4	0.586 [0.395, 0.729]
Item 5	0.464 [0.242, 0.639]
Item 6	0.674 [0.511, 0.790]
Item 7	0.670 [0.506, 0.787]
Item 8	0.588 [0.399, 0.729]
Item 9	0.666 [0.501, 0.784]
Item 10	0.655 [0.486, 0.777]

Note. ICCs are item-level intraclass correlations computed between T1 and T2 with a 1-week interval. All *p*s < .001.

Table S11

Demographic Distribution for Study 6

Demographic variables	Distribution
Age	M_{age} = 39.91, SD_{age} = 13.72, range = 18–78
Sex	Female (*n* = 87), Male (*n* = 73)
Income (‘What is your annual income?’)	≤ USD 15,000 (*n* = 21), USD 15,001–25,000 (*n* = 15), USD 25,000–35,000 (*n* = 13), USD 35,001–50,000 (*n* = 18), USD 50,001–75,000 (*n* = 32), USD 75,001–100,000 (*n* = 31), USD 100,001–150,000 (*n* = 14), ≥ USD 150,001 (*n* = 16)
Subjective Social Status (“The ladder shown represents where people stand in their communities. At the top of the ladder are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. Please place an ‘X’ on the rung that best represents where you think you stand on the ladder (1 = Worst to 10 = Best)”) 	*M* = 5.53, *SD* = 1.60, range = 2–9
Education (‘What is your highest education level?’)	High school (*n* = 37), Associate's degree (*n* = 19), Bachelor's degree (*n* = 73), Master's degree (*n* = 20), Doctorate degree (*n* = 7), Others (*n* = 4)
Race	Caucasian White (*n* = 104), African American (*n* = 17), Hispanic (*n* = 10), Asian (*n* = 20), Mixed (*n* = 9)
Nationality	American (*n* = 144), Non-American (*n* = 16)