SANDIA REPORT
SAND2023-14272
Printed December 2023

Reading Between the Lines:
Measuring the Effects of Linguistic-Based Indicators of Deception on Experts' Identification and Categorization of Disinformation

Matthew B. Windsor, Danielle S. Dickson, Benjamin F. Emery, and Thushara Gunda

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods/

ABSTRACT

There is currently very limited research into how experts analyze and assess potentially fraudulent content in their expertise areas, and most research within the disinformation space involves very limited text samples (e.g., news headlines). The overarching goal of the present study was to explore how an individual's psychological profile and the linguistic features in text might influence an expert's ability to discern disinformation/fraudulent content in academic journal articles. At a high level, the current design tasked experts with reading journal articles from their area of expertise and indicating if they thought an article was deceptive or not. Half the articles they read were journal papers that had been retracted due to academic fraud. Demographic and psychological inventory data collected on the participants was combined with performance data to generate insights about individual expert susceptibility to deception. Our data show that our population of experts were unable to reliably detect deception in formal technical writing. Several psychological dimensions such as comfort with uncertainty and intellectual humility may provide some protection against deception. This work informs our understanding of expert susceptibility to potentially fraudulent content within official, technical information and can be used to inform future mitigative efforts and provide a building block for future disinformation work.

ACKNOWLEDGEMENTS

The authors would like to extend our sincere thanks to a large number of individuals who have been instrumental in the success of the project along the way:

- Nicole Murchison for development of the initial project plan and approach
- Danielle Sanchez, Aaron Jones, Brad Robert, Emily Kemp, and Alisa Rogers for their efforts on study design, implementation, analysis, and execution
- Kurtis Shuler, Lawrence Allen, and Marie Tuft for data processing assistance and statistical consultation
- Shanle Longmire-Monford and Kyle Selasky for qualitative data processing and analysis
- Tres Chaires and Damian Lopez for budget management and financial assistance
- Curtis Johnson for mentorship and Project Management
- Jason Morris and Mallory Stites for consultation and project execution feedback

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronym/Term	Definition
ABM	Agent Based Model
BFI	Big 5 Inventory
CIHS	Comprehensive Intellectual Humility Scale
CRT	Cognitive Reflection Test
HSE	Human Subjects Experiment
GLME	Generalized Linear Mixed Effects
LGBT	Lesbian Gay Bisexual and Transgender
LIWC	Linguistic Inquiry and Word Count
NFC	Need for Closure
NLP	Natural Language Processing
POS	Parts-Of-Speech
SEC	Securities and Exchange Commission
SNL	Sandia National Laboratories

This page left blank

1. INTRODUCTION

Disinformation is not a new phenomenon. As long as human beings have been communicating, there have been attempts to intentionally deceive and mislead. The acts of deception and misleading can take place on any scale, from person to person or country to country. While the advent of electronic communication and the global connectivity characteristic of the information age has certainly changed the landscape of disinformation, humans have been struggling with issues of information veracity for millennia. Interestingly, one of the first known instances of "official" disinformation is Octavian's distribution of coins featuring witty quips and lies slandering Marc Antony during their power struggle to take over leadership of Rome following the assassination of Julius Caesar (Kaminska 2017; Posetti & Matthews 2018). Unfortunately, technology has advanced well beyond necessitating the distribution of physical coins to get a deceptively motivated message across, and thus it becomes more prudent than ever to leverage knowledge, research, and technology to best understand how to mitigate the impacts of disinformation.

Disinformation in the intelligence and National Security space has the potential for disastrous consequences. It is therefore imperative to understand how we can minimize or negate the impacts of disinformation at the individual level, avoiding propagation and dissemination as best as possible. Additionally, full analysis of National Security information often requires extensive knowledge, training, and expertise, but it is not well understood how expert populations fare in the face of disinformation compared to non-experts.

This report first offers a conceptual discussion of disinformation as a construct, the impacts of disinformation, linguistic features of disinformation, and research about human susceptibility to disinformation. Following that, we discuss a research effort that sought to combine human behavioral data, linguistic analysis, and modeling efforts to gain a better understanding of disinformation susceptibility among a population of experts analyzing potentially deceptive formal technical writing in their domain of expertise.

1.1. Disinformation versus Misinformation

An important clarifying distinction to make is the difference between misinformation and disinformation. Both terms deal with the sharing of information that is false, untrue, or misleading with the difference coming from the motivation for sharing. Misinformation is the sharing of false information unintentionally and without harmful intent. Disinformation is the intentional sharing of false information with ill intent or to cause harm, distress, or confusion (Fallis 2014). In other words, disinformation comes from the deceptive and disruptive, and misinformation simply comes from the misguided.

Disinformation exists on a continuum and not all disinformation is necessarily evil or damaging in nature (Fallis 2015). Arguments can be made that marketing and advertising campaigns frequently attempt to intentionally mislead individuals to some extent, and certainly this is not the same level of severity as attempting to ruin a person's life or career with lies and slander. The desired result of intentionally sharing false information can take on a variety of forms. It can be done for personal or financial gain, to sew discord and chaos, to disrupt a political landscape, or simply for entertainment.

Thus, while disinformation is often discussed as a singular construct, it is likely that that the characteristics, mechanisms, and effective mitigatory efforts are unique to disinformation of different forms. Additionally, while disinformation is characterized by intentional sharing, misinformation can be equally disruptive. Often misinformation starts as disinformation and then spreads and propagates via misguided individuals who, while potentially falling prey to their own biases or failures of critical thinking while perusing social media, foster no intent to purposefully deceive. Much of the information spread across social media and traditional news sources contains falsehoods and general misinformation, unknown to those circulating the content (for review, see Lewandowsky et al. 2012). Preventing this cascade of disinformation from transitioning to misinformation as early as possible, before it takes root and spreads, is a critical step in preventing the negative impacts of fraudulent information.

1.2. Avenues and Outlets for Disinformation

1.2.1. The Internet and Social Media

As mentioned above, disinformation can occur at almost any scale. Instead of a comprehensive discussion of disinformation impacts starting at a personal, individual scale, this discussion will focus mainly on disinformation impacts at the societal, or national scale. It is certainly possible to review the historical impacts of disinformation and disruptive information going as far back as the Roman Empire and the eventual advent of the printing press (see Posetti & Matthews 2018 for a brief history), but a full discussion of how disinformation has evolved and shaped history is beyond the scope of this report.

Currently, the likely most salient outlet and source of disinformation is the internet; specifically, social media. Online platforms have enabled unprecedented information sharing capabilities, and these capabilities have irrevocably shifted the information landscape. As such, the impacts of disinformation have never been felt so acutely. Disinformation is not just limited to the spreading of incorrect information by individuals but can also be coordinated malicious campaigns wherein manipulative, often sensationalist or divisive, content is deliberately planted with the intent to influence behavior and social dynamics (e.g., Broniatowski et al. 2018).

While the internet certainly makes it easier and more accessible, misleading or disruptive information sharing is not new and is not limited to social media; indeed, governments regularly engage in acts of misdirection and misinformation to advance their agendas (e.g., in World War II, information regarding D-Day was strategically withheld). However, state-sponsored disruptive information campaigns are now easier than ever, and disinformation has been used to create confusion and sow discord by manipulating information surrounding the Black Lives Matter movement, vaccines, the 2016 Election, COVID-19, and Lesbian Gay Bisexual and Transgender (LGBT) issues (Jankowicz 2020; Muhammed & Mathew 2022; Shu et al. 2020).

The high-profile nature of these online disinformation campaigns has led to an increased desire to understand what causes individuals to share or otherwise interact with (e.g., liking, commenting) this content, how well people can discern true from false information online, and what mitigations to prevent the dissemination of false information online would be most effective. As such, the research in this domain is fairly new and unresolved – most articles in this domain cited regarding these topics will be published after 2016, for example – and academics across disciplines are approaching this topic from multiple perspectives.

1.2.2. National Security Analysis

Beyond sociocultural disinformation campaigns, adversaries of the United States also can actively seek to mislead intelligence analysts by injecting disinformation into an already oversaturated information environment. Adversarial efforts designed to spread disinformation can pose a direct threat to national security, especially assessments of research and development of technology with potential military and offensive implications. As an example, an analysis of North Korean studies in scientific journals, conducted by the Middlebury Institute of International Studies, suggests Pyongyang may be circumventing sanctions through open research collaborations with other countries (Brumfiel & McMinn 2018). This type of evaluation is difficult and precludes comprehensive assessment due to the volume of scientific research and the expertise needed to properly evaluate papers for subtleties that may indicate dual-use research. Reviewing potential dual-use research offers a particularly precarious example of deception detection because, by its very nature, the research has multiple uses. Conclusively inferring that the more military or sinister application is being pursued is a non-trivial task.

Accurately assessing data sources is a crucial role of expert analysts in high-consequence analysis and decision making. Effective analysis and decision making not only requires specialized knowledge and expertise, but also a kind of meta-knowledge about human characteristics, biases, and cognitive shortcomings. The literature makes explicit that it is challenging for people to identify false or misleading messages based on content alone (Lewandowsky et al. 2012; Nemr & Gangware 2019; Shu et al. 2020) and so it is imperative to not only understand all that we can about the features of deceptive content, but also the characteristics of individuals that make them more or less susceptible to disinformation, hopefully with strategies for mitigation.

1.2.3. Financial Reports & SEC Filings

In addition to efforts interested in detecting disinformation campaigns where there are implications for national security, there has historically been interest in the detection of intentional deceit in business reports. Although the motivations in this domain are largely financial – e.g., the context of this deception is typically to retain shareholder confidence and/or seek out investors by composing persuasive reports about company performance - fraud detection has increasingly been utilizing similar techniques to other efforts to detect and study deception (e.g., Burgoon et al. 2015; Humpherys et al. 2011; Markowitz et al. 2021). Thus, findings from this area have been informative for other domains of deception detection.

In particular, the use of specific linguistic patterns has been studied and characterized by utilizing filings for publicly traded companies made available by the Securities and Exchange Commission (SEC). If the SEC detects fraud, those records are also publicly available – thus creating a corpus of fraudulent and non-fraudulent financial reports amenable to academic study (e.g., Humpherys 2009). One of the most prominent findings from this domain is that written reports from underperforming companies will use deliberately obscuring language (i.e., more complex words and sentences) such that their readability is lowered (Ajina et al. 2016; de Souza et al. 2019; Li 2008; c.f., Lo et al. 2017). Obfuscation is not what typically comes to mind with disinformation campaigns, but preventing information from being spread even while purporting to inform the public, a type of omission through distraction, is well within the scope of this issue (Fallis 2014).

1.2.4. Academic Writing

Deception and disinformation within academic writing offers another potentially disastrous example beyond the case of dual-use research discussed above. Fraudulent science, fabricated data, and research misbehavior not only threaten scientific integrity and public trust in scientific advancement but can also interfere with scientific progress. In an increasingly competitive academic environment with an increased emphasis placed on volume of publications, it is no doubt that academic fraud is increasingly prevalent (Parkinson & Wykes 2023). Editors and reviewers of journals are experts in their field, but they are often one of the only lines of defense against academic misbehavior and often serve in that role on top of their other duties as an academic.

So called 'paper mills' also offer a challenge to upholding the integrity of scientific publishing. 'Paper mills' are profit-oriented businesses that submit manufactured manuscripts, often using fabricated and manipulated data/imagery, on behalf of authors in order to generate an easy publication (COPE 2020). These papers are designed to appear legitimate, and require dutiful, overt effort to uncover. This has become particularly problematic for publications coming out of China, where a growing research sector requiring significant publication volume for career advancement motivates individuals to produce fraudulent work to get ahead (Olcott et al. 2023).

Responsible parties, including editorial boards of journals, university research integrity or ethics offices, and academics themselves, attempt to combat this in some ways by retracting papers that are found to have not met their standards. Retractions can occur for a number of reasons beyond fraud (e.g. ethics violations or data reanalysis), but misconduct accounts for a majority of retractions (Fang et al. 2012). While the motivations for fraudulent science are likely often personal and selfish in nature as opposed to nefarious, there is still a clear intent to deceive. Retracted academic papers thus offer an interesting analog and similar challenge as to the analysis of research with potential dual-use implications: both require specialized technical expertise, and both require "reading between the lines" to determine if the author is being deceptive in some way.

Unfortunately, there is a relative paucity of research effectively characterizing the features, indicators, and regularities of academic disinformation. Markowitz and colleagues have attempted to provide a characterization of the linguistic content of fraudulent science (detailed discussion in 1.3.2) both for the case of a single, prolific fraudulent scientist (Markowitz & Hancock 2014) and a varied body of 253 papers retracted for fraudulent data (Markowitz & Hancock 2016). There exists a gap and opportunity to further explore the features of deceptive academic writing, and how to characterize disinformation in formal technical contexts.

1.3. Previous Research on Disinformation

1.3.1. Deception Detection

1.3.1.1. Lie and Deception Detection

While there is a lot of lore, conjecture, and hearsay about effective strategies people can utilize to detect deception, especially within certain professions and in popular media, the research paints a different picture. People generally perform just a little above chance at detecting truth from lies, at least in the broader deception literature (Bond & DePaulo 2006). Even individuals who hold occupations that regularly involve deception detection (e.g., law enforcement, auditors, and members of the judicial system) show no better deception detection than individuals who do not hold such an occupation (Bond & DePaulo 2006; DePaulo & Pfiefer 1986). Indeed, for deception conveyed through speech, only minimal effects of individual differences have been identified, and a more pervasive effect is related to one's general inclination to believe others are being truthful in general than necessarily in their ability to accurately distinguish truth from lies (Bond & DePaulo 2008).

Reliance on verbal or nonverbal cues alone has not improved accuracy outcomes over the years, and alternate strategies that do lead to improvements in deception detection rely on access to the deceiver (e.g., questioning them directly) or contextual information about the situation that would not be available through social media or news sources where the source is purposefully obscured (Levine, 2015). Training in deception detection (Smith 2001) has found it to be only effective in a subset of assessors or outright ineffective in other cases (e.g., Curtis 2021; Zloteanu et al. 2021).

Part of the poor showing in cue-dependent studies might be attributed to the human tendency towards rating others as truthful (truth bias; Levine et al. 1999; Burgoon, Blair, & Strom, 2008; Hartwig & Bond, 2014) and design choices with a 50-50 truth/lie ratio of content to be judged, combined with analysis approaches that do not separately analyze detection of truths from detection of lies (Levine, 2010). In other words, people may be disproportionately poor at detecting lies due to their bias towards assuming most people are telling the truth – a concern given the context of disinformation's prevalence in news and social media. In domains where lying is perceived to be more frequent, such as law enforcement, this bias can reverse, with an overestimation of dishonesty rather than an assumption of truth (Meissner & Kassin 2002). In either case, human accuracy in deception detection is deficient.

Self-assessments of lie detection also reflect a discomfiting decoupling of confidence ratings and accuracy, suggesting low meta-cognitive awareness of one's own ability to detect a lie (DePaulo & Pfiefer 1986; DePaulo et al. 1997). In other words, individuals often express high confidence in their ability to detect a lie but are not actually able to successfully do so. It has been shown that prior exposure to an idea leads people to perceive that idea as more truthful later, called the illusory truth effect (first characterized by Hasher 1977; for more recent meta-analysis see Dechêne 2010). Although this effect can be reduced if the false information is preceded by a warning alerting people to potential falsehoods (e.g., Jalbert, Newman & Schwarz 2020; Lewandowsky & Van Der Linden 2021), warnings *after* exposure vary in efficacy (Greene et al. 1982).

In general, correcting misinformation after it has been disseminated is difficult (e.g., De Keersmaecker & Roets, 2017; see Lewandowsky et al. 2012 for review). The persistence of misinformation poses a particular challenge in the emotionally laden domain of vaccine hesitancy, where well-intended educational efforts may even have backfiring effects (Nyhan, Reifler, Richey, & Freed, 2014; Nyhan & Reifler 2015). Taken together, individuals are susceptible to believing an idea is true just because they read it before, are at risk of maintaining false beliefs even after being told the original source conveyed misinformation (*if* a person can remember the source at all, see Johnson & Hashtroudi 1993 for more on "source monitoring"), and are unable to rely on their own sense of certainty regarding the veracity of what they encounter.

1.3.1.2. Individual Susceptibility Factors for Disinformation

Are there certain features of some individuals that lead them to be more or less susceptible to disinformation? Most attempts to answer that question have examined the propensity of individuals to engage with content on social media (i.e., measuring an individuals' likelihood of spreading of disinformation through a network) or have asked individuals to perform evaluations of headlines or excerpts from news for trustworthiness (i.e., directly measuring an individual's ability to evaluate information credibility). Research to isolate protective traits and risk factors for susceptibility have suggested several candidate dimensions, including cognitive ability, the ability to reflect, intellectual humility, need for closure (dissatisfaction with ambiguity), and conscientiousness or other personality factors, among others (e.g., Bowes & Tasimi 2022, Buchanan 2020, 2021, Buchanan &

Benson 2019, Calvillo et al. 2021, Evans et al. 2020, Koetke et al. 2022; Marchlewska et al. 2017; Mosleh et al. 2021, Newman et al. 2020, Pennycook & Rand 2019, 2020, & Zrnec et al. 2022). Due to the heterogeneity of methodological approaches and measures, trying to synthesize and generalize for a comprehensive meta-analysis is not yet possible (see Bryanov & Vziatysheva 2021 for thorough discussion, but for a brief summary of documented effects in political disinformation, see Sindermann et al. 2020). Finally, it is not clear how well conclusions drawn from identification of false news translate to susceptibility to deception in more formal writing about specialized topics, so generalizations of research in this domain may not be applicable to the current study. Below, the most likely applicable candidate factors are reviewed.

Cognitive ability has been shown to be related to susceptibility to disinformation, with those higher in cognitive ability exhibiting less susceptibility (Sindermann 2021), but cognitive ability is not a unitary construct. Crystallized intelligence refers to knowledge and information that an individual has gained through education and experience whereas fluid intelligence refers to more domain-general cognitive ability such as logic, problem solving, and pattern recognition (Brown 2016). Each facet of intelligence is important, and previous research (Sindermann 2021) suggests that each dimension could have a differential impact to deception detection in terms of learned, domain expertise versus general critical thinking ability. Overall, raw intelligence alone does not seem satisfying at capturing the types of critical thinking skills or tendencies that might protect against disinformation. Correspondingly, more of the research about individual susceptibility to disinformation has focused on utilizing measures that attempt to assess different, potentially protective, aspects of an individual's style of thinking.

One such measure, the Cognitive Reflection Test (CRT) is designed to capture an individual's ability to reject an initial, seemingly obvious (incorrect) solution in favor of the correct solution (Frederick 2005). The CRT assessment is a short, three item assessment that takes the form of simple word problems that appear straightforward on the surface but are specifically designed to require additional reflection beyond the obvious to reach the truly correct solution. The ability to overcome an initial prepotent response is thought to reflect the ability to reject "miserly" thinking that relies too heavily on heuristics and not enough on effortful, deliberate thought. It is also thought to be a uniquely compelling measure of cognitive style in that it is a performance measure as opposed to a self-report measure. In the current context, not accepting information at face value and thinking effortfully, and deliberately, about research results and claims is likely important for those individuals attempting to uncover or detect deceptive content. Indeed, higher scores on the CRT have been associated with lower susceptibility to disinformation (Pennycook & Rand 2019; Pennycook et al. 2020).

Another facet of potential susceptibility to deception and/or disinformation has been captured by assessments of intellectual humility. Intellectual humility refers to the extent to which an individual is aware of, and unthreatened by, the possibility that their knowledge and viewpoints might be imperfect, which may manifest as an openness to changing their mind and/or not being defensive during intellectual disagreements (Krumrei-Mancuso & Rouse 2016). The most well-known measure of this construct, the Comprehensive Intellectual Humility Scale (CIHS), is a multifaceted construct that measures several aspects of intellectual humility, including the independence of an individual's intellect from their ego, their openness to revising their viewpoint, their respect for others' viewpoints, and their *lack* of intellectual overconfidence (i.e., appropriately calibrated intellectual confidence). Scoring high on the intellectual humility scale does not imply a lack of confidence in one's own beliefs or knowledge, but rather implies that the individual has an appropriate grasp on the possibility that their knowledge may be fallible and require adjustment. Intellectual humility, along with the cognitive ability measures listed above, could fall under a broader umbrella of critical thinking skills, with intellectual humility reflecting the ability, or willingness, to revise your opinion and update based on new information. As with cognitive ability, higher intellectual humility scores have been associated with lower susceptibility to disinformation (Bowes & Tasimi 2022; Koetke et al. 2021), though whether that will extend to evaluation of academic disinformation is yet to be determined.

A third individual cognition style that is potentially related to susceptibility to disinformation is captured by the Need for Closure (NFC) scale (Roets & Van Hiel 2011; Webster & Kruglanski 1994). In general, individuals who score high in NFC tend to be uncomfortable with ambiguity, disorder, and a lack of structure. Those with higher scores in NFC tend to prefer to reach quick, unambiguous decisions and dislike having their current knowledge and understanding challenged by conflicting or complicating information. The evaluation of technical, scientific information often requires dealing with a certain amount of ambiguity and nuance, and an individual's NFC may influence how comfortable with or accepting they are of this style of writing. This desire to have definitive knowledge as opposed to uncertainty and having a greater need for clean answers has been associated with greater propensity to engage in conspiratorial thinking (Marchlewska, Cichocka, & Kossowska, 2018), though whether that can be extended to susceptibility to disinformation more generally has not been empirically demonstrated. However, it is possible that a high NFC might discourage thinking too in depth about potentially problematic or inconsistent information, leading to lower ability to detect disinformation in formal technical writing.

Finally, one of the most well-researched frameworks for evaluating personality is the Big 5 Inventory (BFI). The BFI attempts to fully characterize an individual on five scales: (1) Openness to Experience, (2) Conscientiousness, (3) Extraversion, (4) Agreeableness, and (5) Neuroticism (Caspi et al. 2005; Costa & McCrae 1999; Roberts & Yoon 2022). Due to its prominence in the personality literature, there have been many attempts to connect Big 5 traits to susceptibility to disinformation (e.g., Buchanan 2020, 2021; Buchanan & Benson 2019; Calvillo et al. 2021; Evans et al. 2020; Sindermann et al. 2021; Wall et al. 2019; Wolverton & Stevens 2020; Zrnec et al. 2022). Because the measures and designs used vary across this research area, it is hard to generalize the outcomes, but multiple studies have related conscientiousness to lower susceptibility to disinformation (Buchanan 2021; Calvillo et al. 2021; Zrnec et al. 2022). Conscientiousness is also potentially the most compelling factor for the evaluation of academic writing, wherein those scoring lower in conscientiousness might not dedicate thorough attention to studying the quality of the documents.

The majority of this academic research on disinformation has centered on bite-sized information that can be evaluated rapidly to generate many trials, which is optimal for experimental design (to obtain sufficient statistical power and allow for as many performance opportunities as possible). This means participants typically evaluate headlines or paragraph excerpts from larger documents (see Bryanov & Vziatysheva 2021 for a review). However, this leaves a substantial gap in the field, as disinformation may also be disseminated in more official contexts and in longer formats than catchy headlines and short news articles. Additionally, it is possible that the evaluation of "long form" disinformation content draws on a differential set of abilities and psychological characteristics than making judgements of headlines or short paragraphs. The processes used to evaluate deceptive information over time, with evidence and suspicion either building over time or requiring the integration of many pieces of disparate information, is likely a unique exercise.

The present study seeks to examine the evaluation of "long form" disinformation where the goal of the deception is not to catch attention and convey false information quickly, but rather to convincingly present deceptive technical information over the entirety of a formal report. Thus, some of the previously identified candidate factors may not apply, although we will still utilize the battery of assessments identified above to both characterize our population of experts and explore potential links between each measure and the expert readers' ability to detect dishonesty and/or deception in academic writing (see 2.3.3).

1.3.2. Linguistic Features of Disinformation

Given the challenges with human detection of deception, both in person and in written transcripts, it is appealing to have unbiased and objective sources of additional information to guide these judgements (though see also, Heydon 2008, who offers challenges to the field of lie detection in general). Stakeholders with an interest in credibility judgments and the characterization of linguistic regularities have taken advantage of computing advances to analyze text for systemic linguistic cues of deception.

This began with analysis of written transcripts of in-person interactions and was extended to include more experimentally controlled scenarios where the author is instructed to lie. Here, the findings were at least partially consistent with what had been seen in written transcripts of conversational deception. For example, these experimentally motivated liars similarly expressed more negative emotional affect and fewer first-person pronouns than did truth tellers (Hauch et al. 2015; Newman et al. 2003).

Before going into detail about the idiosyncrasies of linguistic markers of deception, it is valuable to describe some of the high-level conceptual features that tend to characterize deceptive writing. While it is doubtful that these markers will be so pronounced as to be detectable through casual reading, it is still worthwhile to understand the logic and psychology of the underlying linguistic cues of deception. In general, indicators of deceptive writing tend to group into a few broad clusters. The first are indicators that the author is distancing themselves or dissociating from the writing and content conveyed therein (Knapp et al. 1974, reviewed in Knapp & Comaden 1979). Usually, this manifests as language that avoids personal statements, personal engagement, or ownership. With respect to parts-of-speech (POS), this usually means the avoidance of first-person pronouns (e.g., I, me, my) (Hauch et al. 2015; Newman et al. 2003).

A second set of indicators are thought to stem from feelings of guilt, negativity, or defensiveness. This is usually reflected in higher rates of negative emotion words and language of more negative valence in deceptive than truthful communications (Newman et al. 2003). Finally, there are linguistic indicators that demonstrate a lack of specificity and a need to obfuscate. At first glance, lack of specificity and obfuscation may not seem conceptually related, but their relatedness stems from the fabricated nature of false information. It is often too difficult to be convincingly concrete with information that is completely made up. As such, deceptive writing is often more vague, abstract, convoluted, imprecise, and tentative than non-deceptive writing. Again, this can be seen in the subtleties of POS analysis, with deceptive writing showing lower rates of article, preposition, quantifier, and adjective use. This, coupled with longer sentences and high rates of jargon, results in effectively less concrete and less readable text (Pennebaker et al. 2003).

This combination of computational approaches to text analysis coupled with psychological analyses of probable causes provides a unique way of leveraging advances in computing to develop testable hypotheses and frameworks for the psychological study of deception in written technical text.

Relatively few efforts have been made to characterize the specific linguistic cues indicative of deception in academic writing, and what does exist is largely from the work of one researcher (Markowitz & Hancock 2014, 2016; Markowitz et al. 2014). This work utilized retracted articles and compared them to non-retracted articles to identify linguistic patterns of deception in academic fraud, and in one case examined the writing style predatory journals themselves (Markowitz et al. 2016). As in the previously reviewed literature on deception, academic fraud was found to have more obfuscation (Markowitz et al. 2014; Markowitz & Hancock 2016) and fewer details (Markowitz & Hancock 2014). Interestingly, the obfuscation in academic writing seems to mirror the previously describing findings in the domain of financial fraud (see section 1.2.3). However, more negative valence of language was not consistently demonstrated in fraudulent academic work, likely reflecting that technical writing does not typically include emotional expression. This failure to replicate is relevant for the present work as well. Much of the prior literature that characterized common linguistic markers of deception utilized informal written narratives outside of professional contexts (e.g., emails and social media posts, or written transcripts of speech), rather than formal technical writing. Although the present study will take an inclusive approach to examining previously identified linguistic features of deception, many of these features are likely to be similar across retracted and non-retracted articles as these features may be relatively sparse in formal technical writing overall.

1.3.3. Disinformation in "Official" Outlets

As previously alluded to, there is concern that the linguistic cues highlighted in natural or pseudo-natural interactive contexts – e.g., lack of personal pronouns, emotional affect – may not be informative for deception detection in publications due to the impersonal and unbiased nature of the preferred academic writing style. As such, linguistic cues independently uncovered in another domain that share more commonalities with academic writing, corporate financial reporting, have been of interest. Similar to academic writing, financial reports adopt a professional tone and report summary outcomes as well as interpretations of those outcomes. These reports are not always written honestly (1.2.3), and some are later exposed as fraudulent (discoveries of which are publicly available by the SEC at https://www.sec.gov/edgar). The goals of the fraudulent behavior across these domains are similar: amplify results, understate poor outcomes, and obfuscate the writing to misdirect attention from dishonest fabrications or mishandling of data. The findings from the small literature on linguistic cues in academic fraud, and the larger literature of language use in corporate fraud, are largely compatible, especially with respect to obfuscation (e.g., Ajina et al. 2016; de Souza et al. 2019; Li 2008; Markowitz & Hancock 2016).

One of the initial motivations for the study was an interest in exploring analysts' ability to review and detect disinformation or misleading content in official contexts, specifically scientific reports covering research with potential dual-use implications. In this context, if a nation state or researcher is pursuing research for one (likely more militarily inclined) reason, but claiming the research is being done for another more mundane reason, then there may be subtle indications in the writing of this deception and obfuscation of true intent. It is, however, quite difficult to identify a set of research articles that have been proven to definitively feature research being conducted for a hidden, dual-use purpose. As a result, it was necessary to identify an analog set of stimuli that could approximate the features of interest: scientific in nature, requiring expert knowledge, published through an "official" outlet, and ground-truth knowledge of some deception or wrong-doing in production of the content. For the purposes of this study, journal articles that had been retracted due to data falsification, fabrication, or fraud were determined to be a suitable proxy.

1.3.4. Expert Evaluation

There is also little research investigating disinformation and deceptive content evaluation by experts. Both in academic contexts and the evaluation of high consequence National Security information, specialized training, expertise, and knowledge is required to effectively assess information. This requires the integration of domain knowledge, methodological expertise, data synthesis, and research implications to be done effectively. And while Zrnec et al 2022 did examine expertise and report an influence of domain experience on fake news discernment, much of the research on deception detection does not explicitly address expert populations. Most of the existing research attempts instead to characterize psychological characteristics and individual performance of a generally representative population. In order to better understand disinformation detection in contexts most applicable to analysis of information with National Security implications, the current study specifically examines expert populations reviewing information within their domain of expertise.

2. METHODS

The present study evaluated human psychological features and then subsequent performance and behavior during a judgement task about potential deception within scientific articles. Subject matter experts were recruited and participated in three sessions of data collection across separate days and sessions. Descriptions of the participants, materials, individual characterization measures, and procedure are captured below. These are partially duplicated from another SAND report that utilized this data (SAND2023-08981, Emery et al. 2023).

2.1. Participants

Human participants consisted of 23 subject matter experts (i.e., those with advanced knowledge in specific fields) at Sandia National Laboratories (SNL). For the present purpose, we defined experts as individuals who have an advanced degree (Masters degree or above) in one of the predefined domains, or have worked in that area for at least 5 years. Participants were drawn from four domains of expertise: biology (5), chemistry (8), computer science (5), and materials science (5).

Participants were recruited through advertisements in the laboratory daily news, department emails, and flyers distributed to the SNL population and were compensated for their time (6 hours total) through a project and task number for standard work hours at their salary rate. All participants consented to participate in accordance with SNL's Human Studies Board.

2.2. Materials

A total of 16 articles, four per domain, were used in the study. Half (deceptive) were collected from the Retraction Watch Database (Oransky & Marcus 2023l; The Retraction Watch Database, 2022) by domain, and the other half (non-deceptive) were topic-matched, collected from Google Scholar, and checked for absence of retraction history.

The Retraction Watch database includes the reason for retraction in the metadata for each retracted article. All deceptive articles selected for the study were chosen based on a retraction due to falsified data, fabricated data, manipulated images, or other overt data manipulation practices. All papers that were retracted for non-content reasons (e.g., ethics violations) were avoided. This was done to maximize the confidence that the retracted articles used in the study are more likely to contain deliberately deceptive content, at least on the part of one or more of the authors. To add a layer of relevance and more closely approximate the kinds of papers an analyst might encounter when reviewing a publication for dual-use motivated deception and align with our initial motivations, the specific topics within each of the four domains of interest were topic matched to align with topics and keywords present in an existing dataset of publications marked as "of concern" for potential military dual-use potential.

Within each domain (Biology, Chemistry, Materials Science, Computer Science), the four articles were grouped into two sets of deceptive and non-deceptive pairs. Participants read one set during one session (session 2) and the other set the following session (session 3). Article deceptiveness was counterbalanced to ensure that effects of deception detection could not be attributed to regularity in the order they were evaluated by participants (e.g. the deceptive article always being read first in a given session). For example, a given human participant in Biology might have read Articles 1 (deceptive) and 2 (non-deceptive) first, and then read Articles 3 (non-deceptive) and 4 (deceptive), while another Biology participant might have read Articles 4 (deceptive) and 3 (non-deceptive) first and then read Articles 2 (non-deceptive) and 1 (deceptive).

To create neutral and de-identified versions of all the articles for participants to review, journal markings, authors and institutional affiliations, retraction watermarks, and external hyperlinks were stripped from documents prior to the experimental sessions.

2.3. Procedure

2.3.1. Session Breakdown / Overall structure

A total of three sessions were used for the experiment. In the first session, which took an hour, participants filled out a demographic and background questionnaire and completed a battery of personality and cognitive assessments. All questionnaire and psychological assessment data was collected using Checkbox Survey software (Checkbox Technology, Inc. 2023). In the second session, which took up to 2.5 hours, participants read and evaluated two scientific articles from their area of expertise and then completed a post-article review assessment for each article.

In the third (final) session, which also took up to 2.5 hours, participants followed the same procedures as the second session with the exception that at the end, participants additionally completed a final post-experiment questionnaire.

2.3.2. Session administration

All sessions were run virtually through Sandia's network. Proctors guided participants remotely via Microsoft Teams, answering questions and administering next steps in the protocol as needed. During the article review portion of the study, participants were allowed to look up supplemental clarifying information online but were asked to constrain their searches to clarifying general information and not look up the article itself.

2.3.3. Psychological Profile Measures

The first session of the experiment focused on collecting data about each individual participant. Data collected about each participant was broadly divided into two categories (1) personal background, demographic, and expertise information and (2) cognitive and personality assessments. Participants first completed a demographic and background questionnaire. This questionnaire asked participants' age, which of the four domains they had expertise in, how many years of experience they had in their field, sub-topics of expertise within their field (open text entry), and their familiarity with the topics of the articles on a 5-point Likert scale ranging from 1 (no experience at all) to 5 (world's pre-eminent expert). We also asked each participant to indicate if they were an "analyst" which was defined as someone who digests information and makes determinations regarding dual-use or deception as part of their daily job.

Additionally, participants completed a series of cognitive and personality assessments. Together, we refer to the personality and cognitive assessments as the "psychological profile" assessments. In all, there were five assessments in our psychological profile assessment:

1. Shipley Intelligence Assessment

2. Comprehensive Intellectual Humility Scale
3. Need for Closure
4. Big 5 Personality Questionnaire
5. Cognitive Reflection Test

Details about each inventory are below.

2.3.3.1. Shipley Cognitive Ability

The Shipley-2 test was used to provide a general assessment of two facets of cognitive ability. The test offers standardized scores assessing cognitive functioning and provides separate assessments that seek to evaluate crystallized and fluid intelligence, which are distinct aspects of human cognitive functioning. The Vocabulary scale consists of 40 items requiring participants to select the closest word to a target word from four options and the Abstraction scale presents 25 sequence/pattern completion items (Kaya et al. 2012; Shipley et al. 2009).

2.3.3.2. Comprehensive Intellectual Humility Scale

The 22-question Comprehensive Intellectual Humility Scale (CIHS) was used to assess Intellectual Humility (Krumrei-Mancuso & Rouse 2016). In additional to an overall measure of intellectual humility, scores were given on four subscales: (1) independence of an individual's intellect from their ego, (2) their openness to revising their viewpoint, (3) their respect for others' viewpoints, and (4) their ***lack*** of intellectual overconfidence. All question prompts were rated by participants on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

2.3.3.3. Need for Closure

To assess individual's Need for Closure, the current study utilized the abridged, 15-item NFC scale (Roets & Van Hiel 2011) which is a shortened, validated version of the full, 42-item NFC scale (Webster & Kruglanski 1994). While the abridged version does not allow for scoring on individual NFC sub-scales as the full version does, it has been shown to be reliable in assessing dispositional NFC overall.

2.3.3.4. Big 5 Inventory

Participants completed the 10-question Big 5 Inventory (BFI), which theoretically captures all elements of an individual's personality via scores across five independent dimensions. Participants viewed 10 statements for which they were required to rate how likely the statement fit a description of themselves (e.g., 'I see myself as someone who is'... 'talkative') on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). While this abbreviated version of the BFI does not provide the more nuanced dimensional sub-scale scores, it has been shown to be relatively reliable and adequate with respect to the main five dimensions when there are substantial time constraints for administration (Rammstedt & John 2007)

2.3.3.5. Cognitive Reflection Test

The Cognitive Reflection Test (Frederick 2005) is a short, three item assessment that takes the form of simple word problems. For example, "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" Participants provided a free response answer to each of the three questions, which were scored as either correct or incorrect.

2.3.4. Participant Tasks

2.3.4.1. Article Review and Annotation

The main task that participants were asked to complete was to review two scientific articles in their area of expertise and then provide their input as to potential indictors of deception. During article review, participants were asked to (1) highlight text that they believe could be associated with deception along with a note explaining their reason for highlighting, (2) provide general comments about anything they found suspect (along with context for any technical details incorporated in their notes), and (3) provide high-level comments on their overall thoughts. Highlights and annotations were all done using the built-in highlighting and annotation features within their document reader and then sent to the proctors following each session.

2.3.4.1.1. Post-article questionnaires

Following review of each article, participants completed a post-article questionnaire. Participants were asked to make a binary determination of if they thought the article was deceptive or not, along with their confidence (on a 1-10 scale with 1 being 0% confident to 10 being 100% confident) for that determination. If they indicated they had found the article to be deceptive, they were asked to indicate from a list of provided reasons why they found the scientific article to be deceptive. The provided reasons included suspicious figures/graphics, omitted information, the presence of inconsistencies, questionable technical rigor within the document, suspicious writing and language use, and/or suspicious data or results. Participants were allowed to select more than one reason for each article. After making their determination participants were asked about sharing. If they indicated that they thought the article was deceptive, they were asked if they would share the article as an example of deceptive content; if they selected that the article was not deceptive, they were asked if they would share the article based on its technical merit. The post-article questionnaire also asked participants to indicate if their review was constrained by time or technical knowledge.

For potential exclusion, participants reported if they had seen the article before (a binary yes/no question). All participants reported having no prior familiarity with the articles.

2.3.4.1.2. End-of-study questionnaire

After completing all individual document reviews, participants answered general subjective questions about their process for evaluating articles. For example, they were asked which section they believed was most important for detecting deception and what indicators of deceptive content they looked for.

2.4. Analysis approach

2.4.1. Descriptive analyses

2.4.1.1. Questionnaires

Data collected from questionnaires was aggregated into summary tables for ease of reference. Apart from factors of experimental interest like years of experience in their domain, a significant portion of information was not factored into inferential analyses of participants' performance as it was intended to provide descriptive context of our population.

2.4.1.2. Article Meta-data

Background meta-data about the articles (e.g., year of publication, number of authors) is aggregated into Table 1. The full listing of articles included in analysis is provided in 0.

Table 1. Stimulus Article Metadata

Citation	Area	Pub. Date	Retract Date	# Author	# Figure	# Ref	# Pages	Word Count
Bertin et al. 2007	Biology	10/23/2007		7	5	31	6	4546
Khan & Cameotra 2013	Biology	10/1/2013	7/9/2014	3	8	63	10	5009
Innocenti et al. 2002	Biology	8/1/2002		8	5	56	10	5518
Mori et al 2000	Biology	4/1/2000	1/1/2011	6	7	38	9	4927
Dey & Airoldi 2008	Chemistry	8/15/2008		2	7	35	7	3821
Guerra et al 2009	Chemistry	5/13/2009	5/15/2011	3	8	32	9	4393
Casciato et al. 2012	Chemistry	8/14/2012	8/7/2020	4	8	42	7	3913
Wang et al. 2003	Chemistry	11/14/2003		8	5	23	7	2389
Fouladi et al. 2021	Computer Science	8/1/2021		5	15	36	15	6724
Saha et al. 2021	Computer Science	4/15/2021	11/30/2021	6	5	72	15	5827
Chen et al. 2019	Computer Science	6/22/2019	1/26/2021	5	14	76	12	7252
Zhou et al. 2021	Computer Science	11/8/2021		5	10	20	6	3885
Ghaffari et al. 2013	Materials Science	8/15/2013	10/7/2016	7	5	42	7	3796
Murali et al. 2013	Materials Science	9/1/2013		8	3	16	5	2080
Hosoyamada et al. 2016	Materials Science	1/8/2016		4	9	10	8	3185
Mahato et al. 2016	Materials Science	5/10/2016	6/8/2017	4	5	53	9	5399

Retracted Article Citations **Bold and Underlined**

2.4.2. Statistical analyses

2.4.2.1. Correlation analyses

The strength of the linear relationship between all the individual psychological profile measures were calculated with Pearson's correlation coefficient *r*, which ranges from -1 for a perfectly negative correlation to +1 for a perfectly positive correlation. The outcomes are presented in Figure 4, with significant correlations indicated by asterisks (*, **, or *** for different levels of significance). This information was used to inform decisions for joint inclusion of factors in the Generalized Linear Mixed Effects (GLME) models described below and provides general insights into the psychological profiles of our participant population.

2.4.2.2. Generalized Linear Mixed Effects Regression Models

Inferential statistical tests were performed to determine if any of the psychological profile measures were predictive of performance on the deception detection task using generalized linear mixed effects (GLME) models. Simply put, these models assess if scoring high (or low) on any specific psychological profile metric predict better or worse ability to discern if an article is deceptive.

A series of GLME logistic regression models were constructed with the binary outcome of correct or incorrect modeled as a dependent variable, and using different combinations of article type (i.e. retracted or not) and other measures (e.g., psychological profile scores) as predictors. More specifically, a base model was generated with a logit link function with task performance (binary correct or incorrect accuracy) as the outcome variable, article status (retracted or presumed honest) as a fixed effect, and individual participant as a random effect. Psychological profile measures were then added into the model individually as additive fixed effects to see if the additional predictor significantly contributes explanatory value to the outcome measure or not. For those models where the additional variable did reach significance, a more complex model including multiple predictors together was generated to test if each variable contributed additional explanatory value or not. All models were generated with the random intercepts for subjects, but no other random slopes or intercepts were included.

2.4.3. Linguistic Analyses

2.4.3.1. Receptiviti Dimension scores

Receptiviti (Receptiviti Inc. 2022) is a commercially available text analysis platform that provides Natural Language Processing analytics on written text. LIWC, and the Receptiviti suite of analytics overall, provides access to a suite of linguistic analysis capabilities providing metrics on dozens of linguistic, cognitive, psychological, personality, emotion, and social dimensions and is used across both industry and academia (Tausczik & Pennebaker 2010). Included in this suite analytics, is the Linguistic Inquiry and Word Count (LIWC) analysis package. LIWC is considered one of the gold standard text analysis capabilities in academia and has been validated and widely used for years, being cited in thousands of published journal articles (Boyd et al. 2022). Through access to their API, the Receptiviti platform was utilized to capture linguistic measures of interest across retracted documents and documents presumed to be free of intentional deception (non-deceptive). The text from each full document (0) was queried with the Receptiviti API, which quantified the contents for a series of measures which are either normed against proprietary datasets (normed measures) or scaled by relative presence of categories within the provided text (dictionary counted measures). These two subtypes of measures are described in more detail below.

Normed measures are baselined against Receptiviti's proprietary datasets, with possible scores ranging from 0 to 100. The proprietary datasets consist of a large corpus of curated written text and allows comparison of input text to text that is intended to be representative of general written text in the world. For example, to help interpret these metrics, a score of 40 for an inputted text sample would imply that 40% of samples in the curated baseline dataset generate scores that are ***less*** than the calculated score of the input sample. In other word, these scores are relative to the scores generated by an external sample of text maintained by Receptiviti.

Dictionary-counted measures are generated by analyzing submitted text one word at a time. As each new word is encountered, if it matches a word that is present in one of the dimension dictionaries, that categorical scale (e.g. "positive emotion words", "certainty") is incrementally updated to reflect the presence of a word from that category. If a word appears in more than one dimension dictionary, then each category is individually incremented. As with the normed measures, this is also a relative measure, but these scales measure the relative presence of categories ***within*** the submitted text itself rather than compared to an external sample. As such, dictionary counted scores are most useful for comparing one piece of input text to another. These dictionary-counted measures range from 0 to 1, with 0 indicating the lack of presence of words from a given category and 1 indicating that every word identified in the text came from that category. Only words that are present in the Receptiviti/LIWC dictionaries contribute to these scores. If a word is not present in the dictionary (i.e., due to being obscure jargon) then it does not count towards scale denominators.

The following table lists the measures of interest derived from Receptiviti and LIWC, including their type (normed or dictionary counted) and whether they were expected to be higher or lower in deceptive relative to non-deceptive writing based on prior literature. Of note, unlike prior work with larger scales of documents, the present study includes only 16 total articles, 8 per document type condition, and therefore reported findings will be more observational than statistically rigorous.

Table 2. Linguistic Cue Predictions (Dictionary Counted Measures)

Dictionary Counted Measures		
Dimension Measure	**Exemplars**	**Evidence from Literature**
Complexity Measures		
Words per sentence	Average words per sentence	More --> Deceptive
Big words	Percent words 7 letters or longer	More --> Deceptive
Dictionary words	Percent words captured by LIWC	Fewer --> Deceptive
Parts of Speech Measures		
1st person singular/plural pronouns	I, me, myself, we, our, us	Fewer --> Deceptive
2nd/3rd person pronouns	you, your, he, she, they, their	More --> Deceptive
Prepositions	to, of, in, for	Fewer --> Deceptive
Articles	a, an, the	Fewer --> Deceptive
Adverbs	so, just, about, there	More --> Deceptive
Conjunctions	and, but, so, as	Fewer --> Deceptive
Common adjectives	more, very, other, new	Fewer --> Deceptive
Conceptual Word Measures		

Dictionary Counted Measures		
Quantifiers	all, one, more, some	Fewer --> Deceptive
Comparisons	greater, best, after	Fewer --> Deceptive
Differentiation	but, not, if, or	Fewer --> Deceptive
All-or-none (absolutist)	all, no, never, always	Fewer --> Deceptive
Causation	how, because, make, why	Fewer --> Deceptive
Tentative	if, or, any, something	More --> Deceptive
Certainty	really, actually, of course, real	Fewer --> Deceptive
(+) Emotion Words	love, nice, sweet	Fewer --> Deceptive
(-) Emotion Words	hurt, ugly, nasty	More --> Deceptive
Abstraction	spirituality, concept, risky, luck	More --> Deceptive
Concreteness	salty, item, person, wooden	Fewer --> Deceptive

Table 3. Linguistic Cue Predictions (Normed Measures)

Normed Measures	
Dimension Measure	**Description**
Personality Dimensions	
Extraversion	“Sociability and social dominance; a tendency to be positive, friendly, and active, seeking out others’ attention and respect.”
Openness	“Openness to new ideas and feelings; interest in art, complex thoughts, emotions, and progressive politics.”
Conscientiousness	“Adherence to order, rules, and duty; involves self-control, a strong work ethic, and a desire for tidiness or organization.”
Neuroticism	“Vulnerability to stress; tendency to experience negative emotions such as sadness, anxiety, and self-consciousness or embarrassment.”
Agreeableness	“Easygoingness and pro-sociality; desire to make others happy, help people, fit in, and be a good or moral person.”
Receptiviti Summary Measures	
Analytical Thinking	Words that suggest formal, logical, and hierarchical thinking
Authenticity	Degree to which author is self-monitoring; measures communication that is personal, honest and unguarded
Emotional Tone	Value < 50 = Negative Tone; Value > 50 = Positive Tone

Note: Big 5 here is a measure of the text content, potentially reflecting properties of the author. Elsewhere in this report Big 5 is also measured in the more traditional sense as a self-reported personality measure.

2.4.4. Agent Based Model

The study of disinformation tends to fall into one of two camps: (1) characterizing features of individuals and how those features influence disinformation detection or (2) modeling and simulation approaches that attempt to replicate and predict the flow and virulence of disinformation through platforms such as social media. There is increased need and appetite for research that attempts to bridge the gap between these two approaches, using insights from human behavioral studies to build better, more efficacious, and ecologically valid models. Agent Based Models (ABMs), which focus on interactions between autonomous, individual elements, offer one such potential avenue for combining these two approaches and have previously been used to study diffusion of disinformation (Kaligotla et al 2022). Through the design of individual agents, agent interactions, and structure of the model, researchers are able to simulate a variety of scenarios and specific ; situations that would be expensive and time consuming to fully account for through human subjects experimentation, if possible at all. An existing gap is that while ABMs are built around agents and their interactions, thoughtful design of more complex individual agents has been given little attention to date. Within the context of simulating human psychological features and decision making, working towards designing agents within an ABM that can more fully represent the complexity of humans is a worthwhile endeavor. This is especially crucial within the disinformation landscape given the importance that individual differences such as personality and cognitive styles have been shown to have on recognition of deception and disinformation (see Bryanov & Vziatysheva 2021 for a review). The current effort attempted to advance efforts to integrate empirical human psychological and behavioral data into the design and parameterization of an ABM, with a specific emphasis on more deliberate and complex agent design for the simulation of disinformation detection. This work builds on research that couples ABMs synergistically with controlled studies to help isolate specific phenomenon (Duffy 2006). Specific details about the ABM design and parameterization can be found in SAND2023-08981 (Emery et al. 2023).

This page left blank

3. RESULTS AND DISCUSSION

3.1. Participant Demographics

Participants were recruited from four domains of expertise. Demographics about the average age and years of experience for each domain and overall can be seen below in Table 4.

Table 4. Participant Demographic Information

Technical Domain	# of Participants	Avg. Age	Avg. Yrs. of Experience
Biology	5	42	11.4
Chemistry	8	37.9	11.4
Computer Science	5	38.6	10.7
Materials Science	5	44.8	20.8
Overall	**23**	**40.4**	**13.3**

3.2. Psychological Profile Measures

Figure 1 through Figure 3 provide frequency density plots showing our participants' distribution of scores on each of the psychological profile measures. These results help to paint a picture of the psychological characteristics of our expert population. The observed ranges, variability, and correlations (see Figure 4) were used to inform the GLME logistic regression models.

Within the BFI, conscientiousness has the strongest evidence as being protective against deception and our observed distribution of conscientiousness scores is relatively high and narrow. This is perhaps unsurprising given our participant population of technical experts at a National Laboratory, yet the narrow range and near ceiling scores make inclusion in a predictive model inefficacious due to restricted range.

NFC scores were in the middle of the possible range, with a fairly narrow distribution, indicating that our population is neither low nor high on their dispositional need for closure.

Overall intellectual humility scores were on the higher end of the possible range, which is also likely unsurprising given our population of conscientious technical experts. Intellectual humility subscale scores for "Openness to Revising One's Viewpoint" and "Respect for Others' Viewpoints" were on the higher end of the possible range, with "Respect for Others' Viewpoints" being at almost ceiling. Subscale scores for "Independence of Intellect and Ego" and "Lack of Intellectual Overconfidence" were more in the middle and with greater ranges.

Our measures for cognitive ability, Shipley and CRT, showed scores on the higher end as expected. Shipley is useful for comparing our participant population to the general population given that the scores are standardized based on general population scores by age. It should be noted that, the range of scores on the Shipley Abstract assessment was greater than that of the overall and Vocabulary subscale, suggesting that this measure of fluid intelligence is more variable within our participant population.

Table 5. Psychological Profile Measures

Psychological Profile Measure Legend	
Name	**Measure**
BFI10.Openness	Big 5 Inventory Openness to Experience
BFI10.Conscientiousness	Big 5 Inventory Conscientiousness
BFI10.Extraversion	Big 5 Inventory Extraversion
BFI0.Agreeableness	Big 5 Inventory Agreeableness
BFI10.Neuroticism	Big 5 Inventory Neuroticism
NFC.Score	Need for Closure
CIHS.Int_Humility	Intellectual Humility (IH) Overall Score
CIHS.Independence	IH - Independence of Intellect and Ego
CIHS.Openness	IH - Openness to Revising One's Viewpoint
CIHS.Respect	IH - Respect for Others' Viewpoints
CIHS.Overconfidence	IH - Lack of Intellectual Overconfidence
CRT.Score	Cognitive Reflection Test
ShipleyTotal	Shipley Intelligence Overall Score
ShipleyAbstract	Shipley Abstract Reasoning Score
ShipleyVocab	Shipley Vocabulary Score

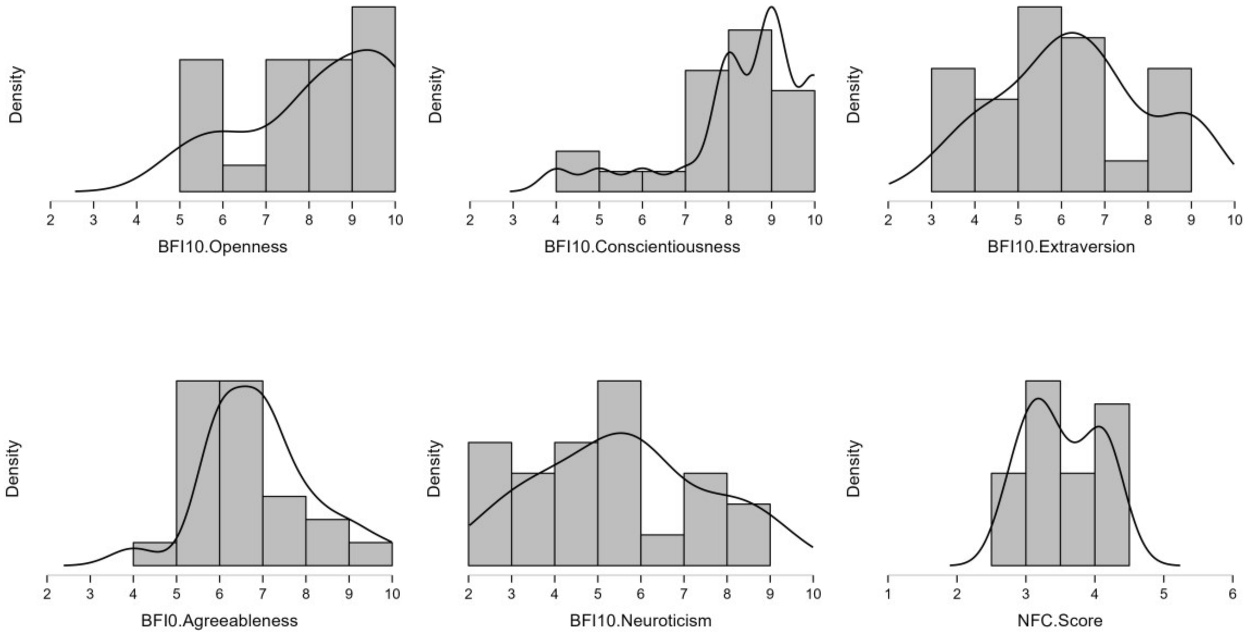


Figure 1. Psychological Profile Score – Frequency/Density Plots (1/3)

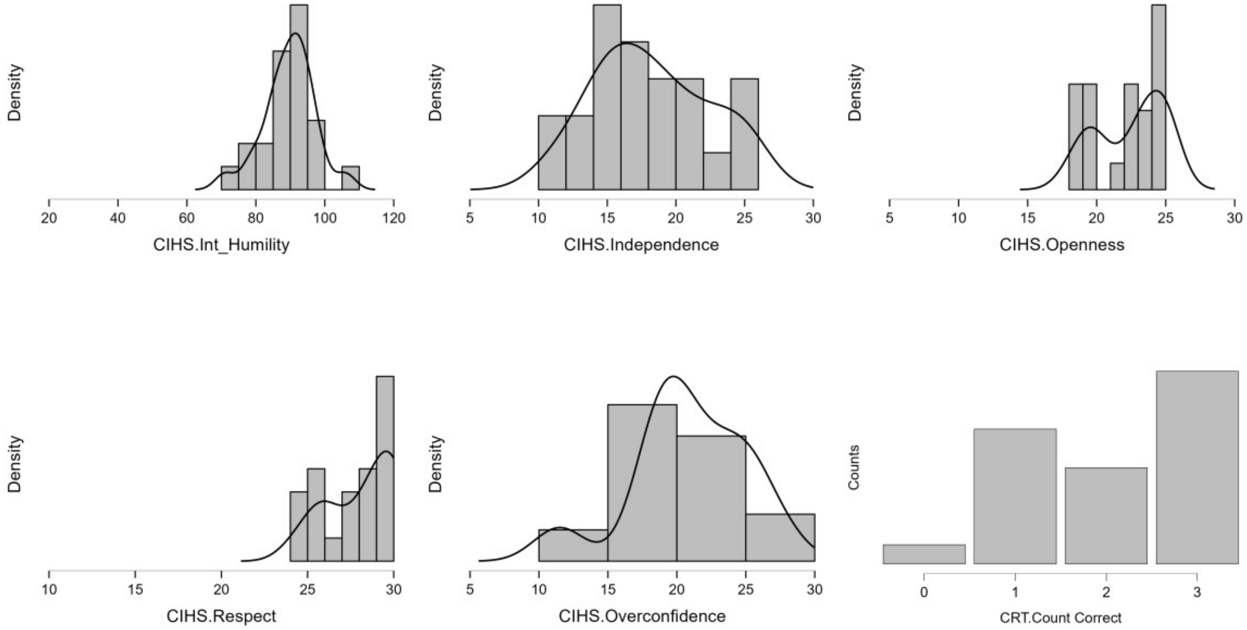


Figure 2. Psychological Profile Score – Frequency/Density Plots (2/3)

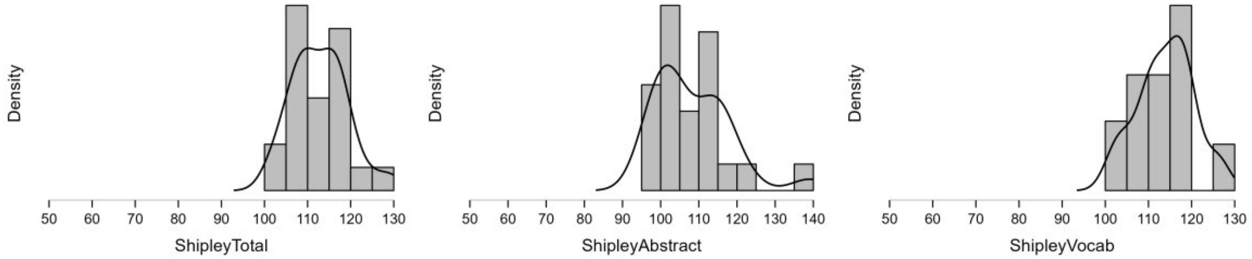
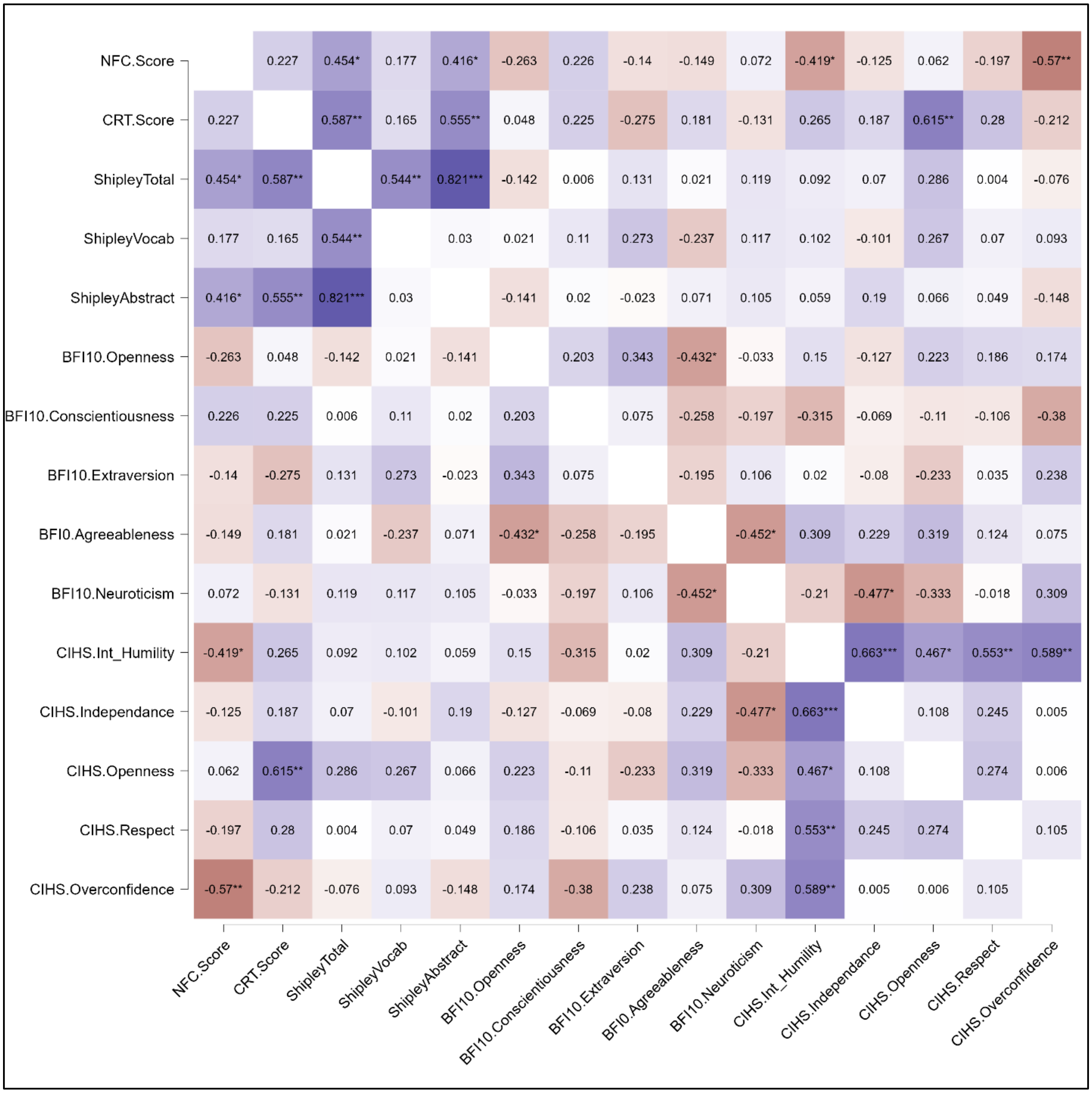


Figure 3. Psychological Profile Score – Frequency/Density Plots (3/3)

3.2.1. Psychological Profile Correlation Matrix

Presented below in Figure 4 is a correlation matrix showing the correlations between each of our psychological profile measures. The color and saturation of each correlation cell indicates the direction and strength of each correlation respectively, with positive correlations in blue/purple and negative correlations in red, with more saturation (darkness) indicating a stronger relationship. All overall and subscale scores were included in the matrix, and thus some of the stronger relationships that immediately stand out due to their saturated color can be ignored due to being subscale score correlations with the respective overall scale.

In general, very few of the psychological profile measures showed significant interrelationships of note. CRT scores displayed a couple interesting relationships, with that measure being positively correlated with CIHS.Openness (*r* = .615) and Shipley Abstract (*r* = .555). This suggests a positive relationship between individuals' tendency to engage in effortful, deliberate thought (CRT) and both fluid intelligence (Shipley Abstract) and openness to other viewpoints (CIHS.Openness). Another interesting trend was the negative correlation between CIHS.Overconfidence ("Lack of Overconfidence") and NFC (*r* = -.57), which suggests that those with appropriately calibrated intellectual confidence expressed a lower need for closure. A final note is that some of these scores are known to be interrelated and tap overlapping dimensions of an individuals' psychological state, and thus care was given in interpretation and model development to account for these known/existing interrelationships.



	NFC.Score	CRT.Score	ShipleyTotal	ShipleyVocab	ShipleyAbstract	BFI10.Openness	BFI10.Conscientiousness	BFI10.Extraversion	BFI0.Agreeableness	BFI10.Neuroticism	CIHS.Int_Humility	CIHS.Independance	CIHS.Openness	CIHS.Respect	CIHS.Overconfidence
NFC.Score		0.227	0.454*	0.177	0.416*	-0.263	0.226	-0.14	-0.149	0.072	-0.419*	-0.125	0.062	-0.197	-0.57**
CRT.Score	0.227		0.587**	0.165	0.555**	0.048	0.225	-0.275	0.181	-0.131	0.265	0.187	0.615**	0.28	-0.212
ShipleyTotal	0.454*	0.587**		0.544**	0.821***	-0.142	0.006	0.131	0.021	0.119	0.092	0.07	0.286	0.004	-0.076
ShipleyVocab	0.177	0.165	0.544**		0.03	0.021	0.11	0.273	-0.237	0.117	0.102	-0.101	0.267	0.07	0.093
ShipleyAbstract	0.416*	0.555**	0.821***	0.03		-0.141	0.02	-0.023	0.071	0.105	0.059	0.19	0.066	0.049	-0.148
BFI10.Openness	-0.263	0.048	-0.142	0.021	-0.141		0.203	0.343	-0.432*	-0.033	0.15	-0.127	0.223	0.186	0.174
BFI10.Conscientiousness	0.226	0.225	0.006	0.11	0.02	0.203		0.075	-0.258	-0.197	-0.315	-0.069	-0.11	-0.106	-0.38
BFI10.Extraversion	-0.14	-0.275	0.131	0.273	-0.023	0.343	0.075		-0.195	0.106	0.02	-0.08	-0.233	0.035	0.238
BFI0.Agreeableness	-0.149	0.181	0.021	-0.237	0.071	-0.432*	-0.258	-0.195		-0.452*	0.309	0.229	0.319	0.124	0.075
BFI10.Neuroticism	0.072	-0.131	0.119	0.117	0.105	-0.033	-0.197	0.106	-0.452*		-0.21	-0.477*	-0.333	-0.018	0.309
CIHS.Int_Humility	-0.419*	0.265	0.092	0.102	0.059	0.15	-0.315	0.02	0.309	-0.21		0.663***	0.467*	0.553**	0.589**
CIHS.Independance	-0.125	0.187	0.07	-0.101	0.19	-0.127	-0.069	-0.08	0.229	-0.477*	0.663***		0.108	0.245	0.005
CIHS.Openness	0.062	0.615**	0.286	0.267	0.066	0.223	-0.11	-0.233	0.319	-0.333	0.467*	0.108		0.274	0.006
CIHS.Respect	-0.197	0.28	0.004	0.07	0.049	0.186	-0.106	0.035	0.124	-0.018	0.553**	0.245	0.274		0.105
CIHS.Overconfidence	-0.57**	-0.212	-0.076	0.093	-0.148	0.174	-0.38	0.238	0.075	0.309	0.589**	0.005	0.006	0.105	

* p < .05, ** p < .01, *** p< .001

NFC: Need for Closure; CRT: Cognitive Reflection Test; BFI: Big 5 Inventory, CIHS: Comprehensive Intellectual Humility Scale

Figure 4. Psychological Profile Measure Correlations.

3.3. Article Review

3.3.1. Deception Detection Performance

3.3.1.1. Classification Performance Results

Table 6 and Table 7 provide the performance and classification results of the study. Each participant read four articles, two of which were articles that had been retracted due to data fraud/fabrication, and then asked to make a binary decision as to whether or not they believed the article was attempting to be deceptive. As a result, each participant had four performance opportunities to correctly classify an article as either deceptive or not. The full confusion matrix of possibilities can be seen in Table 6.

Of particular interest, is the overall accuracy of exactly 50%. Relatively consistent with the literature covering people's deception detection performance, our population of experts were overall at chance at detection of deceptive (retracted) versus not retracted articles. There was a relatively pronounced bias for saying that articles were not deceptive, as can be seen in the False Negative (FN) and True Negative (TN) results. This aligns with the concept of "truth bias" discussed in the Lie and Deception Detection section above. This resulted in lower performance on trials with retracted papers (Accuracy: 39%) versus trials with non-retracted papers (Accuracy: 61%). With respect to incorrect trials, FN were much more prevalent than False Positives (FP), which is potentially problematic for the dissemination of deceptive and fraudulent information. This is also reflected in the Precision and Recall scores of 50% and 39% respectively. While the F1 score is provided, it should be interpreted with caution given the relatively small number of trials and balanced (equal retracted and not retracted) data under consideration.

Additional tables in sections 3.4 and 3.5 provide supporting context for performance using supplementary collected measures from the post-article surveys (e.g., self-reported confidence) and the post-experiment survey.

Table 6. Detection and Classification Confusion Matrix

	TP: True Positive; **FP:** False Positive; **FN:** False Negative; **TN:** True Negative			
		Article Type		
		Retracted	Not Retracted	Total
Participant Response	Deceptive	TP = 18	FP = 18	36
	Not Deceptive	FN = 28	TN = 28	56
Total		46	46	92

Table 7. Additional Detection and Classification Metrics

Measure	Value	Calculation
Accuracy	0.5	(TP + TN) / (P + N)
Precision	0.5	TP / (TP + FP)
Recall (Sensitivity)	0.39	TP / (TP + FN)
F1 Score	0.44	2TP / (2TP + FP + FN)
Specificity	0.61	TN / (FP + TN)
Negative Predictive Value	0.5	TN / (TN + FN)

Measure	Value	Calculation
False Positive Rate	0.39	FP / (FP + TN)
False Discovery Rate	0.5	FP / (FP + TP)
False Negative Rate	0.61	FN / (FN + TP)

3.3.1.2. Performance Modeling

Below are the results of the sequence of models that were run using generalized linear mixed effects logistic regression through the JASP statistics software platform (JASP Team 2023). As outlined above, a base model predicting trial accuracy (correct or incorrect) as a function of article type (retracted versus not retracted) as a fixed effect was built first. Additional models were run individually by adding in each psychological profile measure as an additive fixed effect along with article type to predict trial accuracy. The only random effect included in the model were random intercepts for subjects, with no other random slopes or intercepts included.

Table 8 displays those models for which a psychological profile measure was shown to contribute significant explanatory value to the model. Article type was shown to be a significant factor in all models (all *p's* < .05). In GLME models, each estimate reflects the contribution to the model when controlling for each other predictor. Thus, for each individual model, the effects can be interpreted as the effect of that psychological profile measure on trial accuracy when controlling for article type. Following the development of the individual models, a more complex model was generated that included all of the individually significant psychological profile measures as predictors along with article type. The estimates for that more complex model can be seen at the bottom of Table 8.

As each model was run with a Logit link function, each estimate value represents the change in log-odds associated with a change in that predictor. For a categorical predictor like article type (retracted versus not retracted), this represents the change in log-odds compared to the reference class (not retracted). For continuous predictors, like NFC or CIHS.Overconfidence, this represents the change in log-odds with a one-unit increase in that predictor.

Each log-odds estimate can also be transformed into an odds-ratio (included in the table). Odds-ratios > 1 indicate that a change in that predictor (categorical compared to reference class or 1-unit increment) results in greater odds of the outcome variable (e.g., odds-ratio of 1.55 = 55% increase in odds). Odds-ratios < 1 translate to lower odds (odds-ratio of .85 = 15% decrease in odds).

For the "base" model with only article type included as factor, the coefficient estimate for article type (β = .442; odds-ratio = 1.56) reflects that individuals did not show equivalent performance for retracted vs. non-retracted articles, with individuals being more likely to correctly classify non-retracted articles as such.

Need or Closure (NFC) and the Lack of Intellectual Overconfidence subscale (CIHS.Overconfidence) from the Intellectual Humility assessment provided predictive and explanatory value in the generated models in ways consistent with predictions and the literature. The negative model estimate (β = -.837) for NFC indicates that a higher NFC score will result in a reduced log-odds of correctly classifying an article (odds-ratio = .433; 57% reduced odds). Higher NFC is thought to indicate a more close-minded thinking style and a need for clear, unambiguous answers. This preferred style of thinking could be detrimental to the mindset necessary to detect deception in that, questioning claims, recognizing and exploring inconsistencies, and being open to ambiguity seems to be beneficial to detecting potential deception in academic writing.

It is worth emphasizing again that CIHS.Overconfidence is the ***lack*** of intellectual overconfidence. The positive model estimate (β = .131) for CIHS.Overconfidence suggests that a higher score on that scale increases the log-odds of correctly classifying an article (odds-ratio = 1.14; 14% increased odds). Considering that one's own knowledge could be fallible seems to be an important feature for comprehensive and open-minded evaluation. This suggests that an appropriately calibrated knowledge of one's intellectual strengths and weaknesses can be helpful in evaluating and detecting deceptive technical content.

Shipley Abstract purports to tap into fluid intelligence, which reflects an individual's flexible cognitive ability used in applying logic and problem solving to new situations. The model estimate for the Shipley Abstract measure (β = -.056) is a relatively small effect (odds-ratio = .946; ~5% reduced odds) but it still worth noting as this finding runs counter to the predicted hypothesis. As outlined above, cognitive ability and critical thinking in general has largely been protective against deception and endorsement of fake news. This result should, however, be interpreted with caution due to the relatively limited performance data upon which it was generated.

Finally, when assessing the estimates for the more complex model including Article Type (β = 0.494), NFC (β = 0.017), Shipley Abstract (β = -0.052), and CIHS.Overconfidence (β = 0.121), although Shipley Abstract and CIHS.Overconfidence trend towards significance, only Article Type emerges as a significant predictor, with NFC losing almost all predictive value (odds-ratio = 1.02; 2% increase in odds; *p* > .90).

Overall, when allowing for each participant to have a random intercept in the generated models, Article Type seems to be most stable predictor, with only CIHS.Overconfidence providing somewhat consistent predictive value in the hypothesized direction. While based on somewhat limited performance data, this suggests that a lack of intellectual overconfidence can provide benefits to thoughtful and accurate detection of potential deception in formal academic writing.

Table 8. GLME Logistic Regression Model Results

GLME Logistic Regression – Fixed Effect Model Estimates 23 Participants x 4 Trials = 92 total trials					
Term	**Estimate**	**Odds Ratio**	**Std. Error**	**t statistic**	**p - value**
Accuracy ~ Article Type					
Intercept	-2.40×10^{-10}	.999	0.214	-1.125×10^{-9}	1
Article Type (1)	.442	1.56	0.214	2.07	0.04
Accuracy ~ Article Type + NFC Score					
Intercept	2.955	19.202	1.52	1.944	0.05
Article Type (1)	0.462	1.5872	0.219	2.108	0.035
NFC.Score	-0.837	0.433	0.426	-1.963	0.05
Accuracy ~ Article Type + Shipley Abstract Score					

GLME Logistic Regression – Fixed Effect Model Estimates 23 Participants x 4 Trials = 92 total trials					
Intercept	6.026	414.06	2.696	2.235	0.025
Article Type (1)	0.471	1.602	0.222	2.124	0.034
Shipley Abstract	-0.056	0.946	0.025	-2.235	0.025
Accuracy ~ Article Type + CIHS Lack of Intellectual Overconfidence Score					
Intercept	-2.73	0.065	1.242	-2.198	0.028
Article Type (1)	0.47	1.6	0.221	2.123	0.034
CIHS.Overconfidence	0.131	1.14	0.058	2.241	0.025
Accuracy ~ Article Type + Shipley Abstract + NFC Score + CIHS Lack of Overconfidence					
Intercept	2.993	19.945	3.488	0.858	0.391
CIHS.Overconfidence	0.121	1.129	0.071	1.705	0.088
Shipley Abstract	-0.052	0.949	0.028	-1.86	0.063
NFC.Score	0.017	1.017	0.56	0.031	0.976
Article Type (1)	0.494	1.639	0.228	2.168	0.03

3.3.2. Article Annotations

Across all 23 participants each reading 4 papers, there were 1413 comments made concerning potentially deceptive or concerning content. There were 61.4 comments on average per participant (*SD* = 44.5), 31.6 comments on average for non-retracted papers (*SD* = 23.7) and 29.8 comments on average for retracted papers (*SD* = 21.7). As is clear from the standard deviations, there was quite high variability in our participant pool for how many comments individuals would make. Informal qualitative analysis established a high-level categorization scheme for comments which revealed that comments generally covered concerns related to:

Suggesting deception

- Missing or Omitted Information
  - Missing/suspect data
  - Parameters not listed on samples
  - Missing control groups
  - Missing sections of paper
  - Not listing citations/citations suspect
- Suspicious or Irregular Data
  - Data is too uniform/perfect
  - Use of unusual units
  - Data doesn't match with the paper's topic
- Language and Linguistic Indicators
  - Subjective statements
  - Vague statements
  - Conflicting statements

    - Grammatical errors
- Making leaps and jumping to conclusions on the data

Suggesting legitimacy

- Conscientious Writing
    - Detailed and through sections of paper (e.g., methods, materials, etc.)
    - Images at same magnification
    - Explanations of discrepancies

3.4. Post-Article Questionnaire

Below are a series of tables reporting the data from the questionnaires that each participant completed following their review of each article. These questions were designed to not only assess performance and ability to detect potential deception (covered above) but also gain insights into potential sharing behavior, and determine factors that might influence an individual's ability to adequately perform the task.

Table 9 provides the data, somewhat covered above, for how often participants thought articles were deceptive versus not. Stimuli were balanced such that each participant saw half deceptive and half not deceptive articles. The truth bias (tendency for people to, in the absence of cues to the contrary, tend to rate others as truthful) is abundantly clear.

Table 9. Deception Judgments – Overall Counts

Do you believe the authors were attempting to be deceptive in this article?	
	Count
No	56
Yes	36
Total	92

Table 10 presents the results for how often individuals said they would share the article based on technical merit. It should be noted that this question was conditionally presented to participants if and only if they indicated that the article was not deceptive. This can be confirmed by noting that the overall count of responses to this question matches the count for number of not deceptive judgements. Provided are separate counts for sharing intent based on whether a paper was retracted or not. Interestingly, and somewhat troubling, is that participants were relatively more likely to share retracted than non-retracted articles based on technical merit if they had determined it to be not deceptive. This suggests that if there is not a strong enough signal resulting in an individual to determine a paper to be deceptive, then it is likely to be seen as legitimate and worthy of sharing.

Table 10. Would Share Based on Technical Merit – Not retracted vs. Retracted Articles

Would you be willing to share this article with a customer based on its technical merit?			
	Non-retracted	**Retracted**	**Total**
No	14	9	23
Yes	14	19	33
Total	28	28	56

Table 11 through

Table 13 provide the average confidence ratings for each judgement, broken down in a variety of ways. This question was included to assess if there is perhaps something about the writing in retracted articles that, while not necessarily impacting performance, may influence confidence in a decision. On the whole, this turns out to not be the case. Confidence was roughly the same for trials where participants made accurate versus inaccurate judgments (Table 11), for retracted versus not retracted articles (Table 12), and for all types of classification (Table 13). With confidence being roughly equivalent, and just above the middle of the 1 – 10 range for all situations, it does not appear that article type or detection performance impacts individual confidence in any meaningful way.

Table 11. Average Confidence – Accurate vs. Inaccurate Trials

How confident are you in your categorization of this article as deceptive/not deceptive? (1-10 scale - Low to High)	
	Average
Accurate	6.2
Inaccurate	6.6
Total	6.4

Table 12. Average Confidence – Not Retracted vs. Retracted Articles

How confident are you in your categorization of this article as deceptive/not deceptive? (1-10 scale - Low to High)	
	Average
Non-retracted	6.4
Retracted	6.4
Total	6.4

Table 13. Average Confidence – Classification Confusion Matrix

How confident are you in your categorization of this article as deceptive/not deceptive? (1-10 scale - Low to High)	
	Average
True Negative	6.2
False Positive	6.7
True Positive	6.2
False Negative	6.5
Total	6.4

Table 14 through Table 17 presents the data for questions assessing the potential impact of various factors on participants' ability to perform the task.

The first question asked if an individual felt that their assessment was limited by their technical expertise. Table 14 presents the data for how often participants felt this was the case, separated by trials where participants were correct or incorrect. While it is the case that participants were about twice as likely to say they were limited in their assessment by their technical expertise (and something to consider for methodological improvements going forward), their perceived inadequate technical background did not seem to influence performance, as counts were roughly equivalent for accurate versus inaccurate trials. It is, however, a compelling finding that experts ***felt*** they were constrained by their technical background and could be worth further investigation in the future.

Table 15 shows the results of participants being asked if their assessment was limited by time. Encouragingly, and opposite the technical expertise question trend, participants were about twice as likely to say they were not limited by time, and these rates were the same between accurate and inaccurate trials. Thus, time was likely not a factor in influencing performance.

Table 16 addresses the question of if participants were limited by time a different way, by looking at counts separately for retracted versus not retracted articles. Here, a somewhat more interesting story emerges. For trials where participants did not feel they were limited by time, the count was greater for non-retracted compared to retracted articles. This trend reverses for trials where participants did feel they were limited by time, with a higher count for trials featuring retracted articles. While admittedly a modest trend, this does provide some evidence that there may be something about retracted articles that leads individuals to not feel as if they have enough time to do a thorough review. If deceptive technical writing is in some way, consciously or unconsciously, written in a more complex, vague, and inconsistent way, then this could influence how long it takes an individual to read and feel they have an adequate grasp on the content.

Finally, Table 17 shows the results for the question of if participants felt they were able to review each article in its entirety. Here, a similar trend emerges as to time limitations for accurate versus inaccurate trials, in that there are no meaningful differences between accurate and inaccurate trials in terms of perceived review completion. In addition to participants being about 5 times more likely to feel they were able to completely review the article compared to not, there were no differences between accurate and inaccurate trials. This indicates that participants' ability to completely review each article did not meaningfully impact performance.

Table 14. Limited by Technical Background – Accurate vs. Inaccurate Trials

Was your assessment of this article limited by your technical background?			
	Accurate	**Inaccurate**	**Total**
No	14	13	27
Yes	32	33	65
Total	46	46	92

Table 15. Limited by Time – Accurate vs. Inaccurate Trials

Was your assessment of this article influenced by the time limit?			
	Accurate	**Inaccurate**	**Total**
No	30	30	60
Yes	16	16	32
Total	46	46	92

Table 16. Limited by Time – Not Retracted vs. Retracted Articles

Was your assessment of this article influenced by the time limit?			
	Non-retracted	**Retracted**	**Total**

Was your assessment of this article influenced by the time limit?			
No	34	26	60
Yes	12	20	32
Total	46	46	92

Table 17. Able to Review Entire Article – Accurate vs. Inaccurate Trials

Were you able to review the article in its entirety?			
	Accurate	**Inaccurate**	**Total**
No	7	8	15
Yes	39	38	77
Total	46	46	92

3.5. Post-Experiment Questionnaire

At the end of the experiment, all participants completed a post-experiment questionnaire that sought to provide some insights into how our participant population approached assessment of the articles overall and what they saw as important for determining if an article is deceptive or not. Figure 5 through Figure 8 show each question that was asked, and the relative frequency of responses. While the abstract was seen as the section where most individuals would *begin* their assessment, the tables/figures, methods, and results were seen as the most important sections. This is perhaps unsurprising in that these are the sections where fraudulent data or suspicious methodological practices would be revealed. This is confirmed in Figure 7 where participants reported what kinds of features they would look for when attempting to detect deception, with "inconsistency between reported data and conclusions" and "suspicious figures or misleading figures/graphics" having the highest frequency. It is encouraging that, when asked when they would reach out for assistance, "inadequate technical background" and "potential mission impact" were the two most frequent reasons reported. This is in line with the relatively high intellectual humility scores observed, and demonstrates a respect for the importance of our mission operations and National Security.

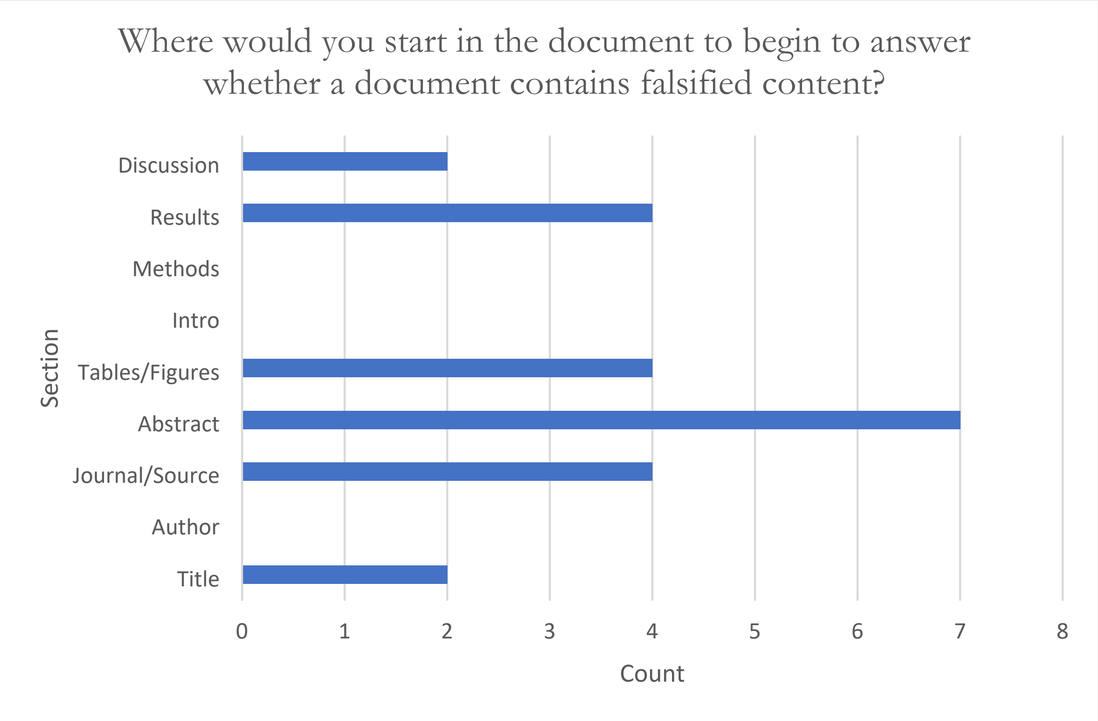


Figure 5. Review Strategy – Where to Start Review?

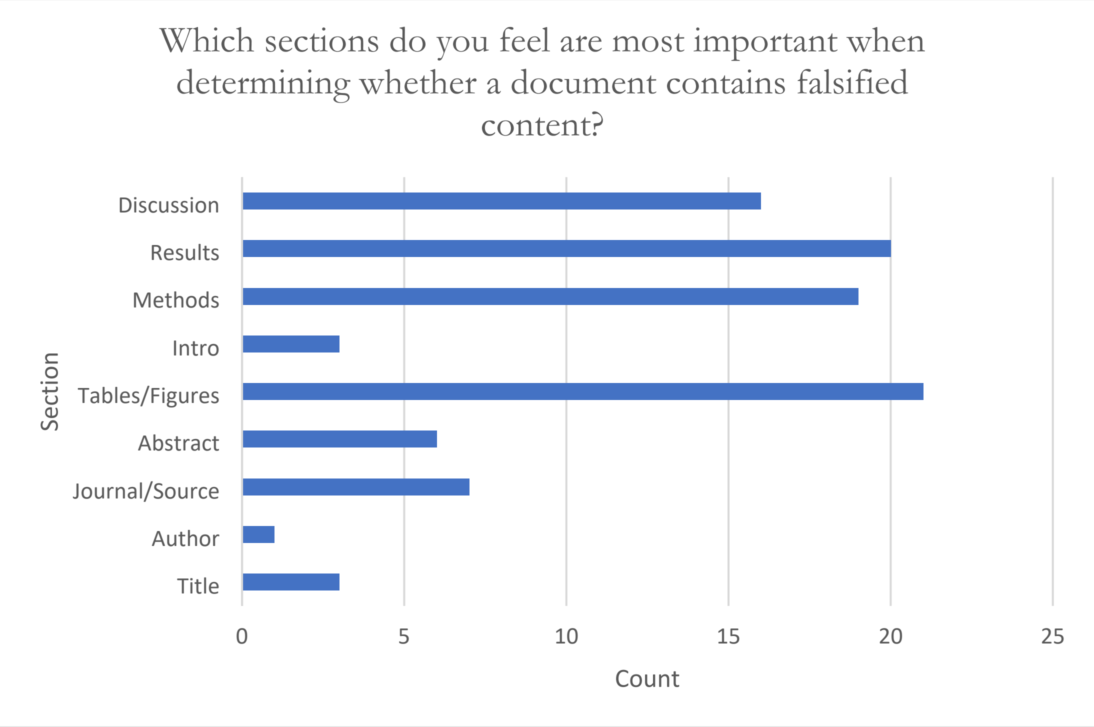


Figure 6. Review Strategy – Most Important Paper Sections

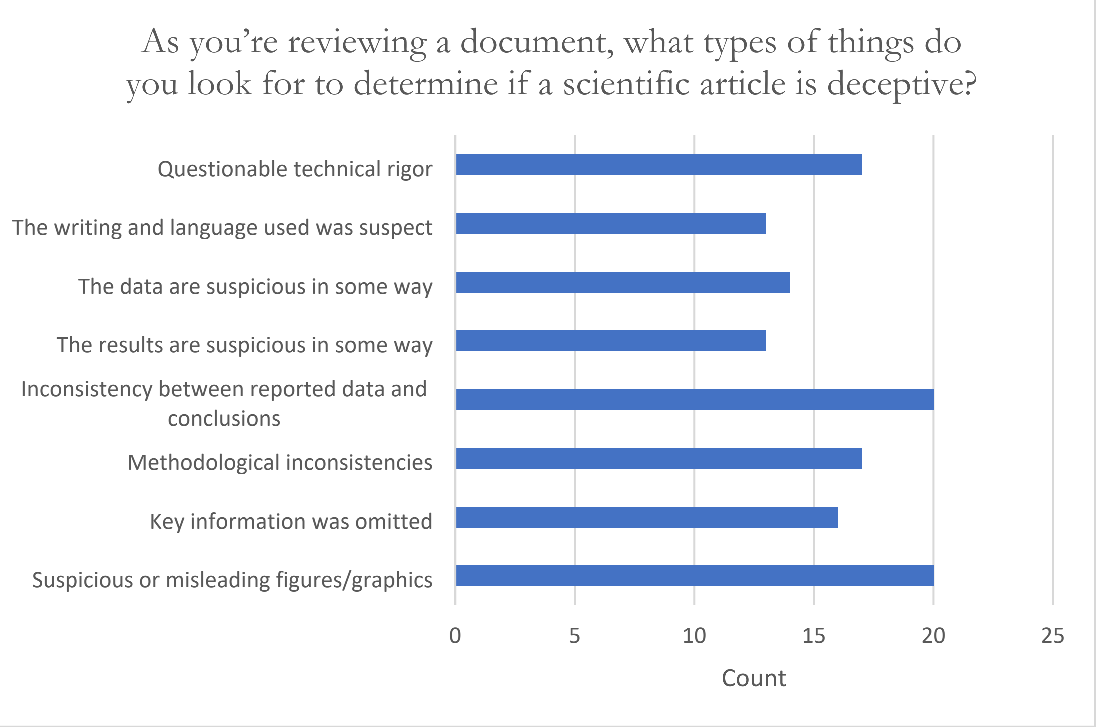


Figure 7. Review Strategy – What to Look For?

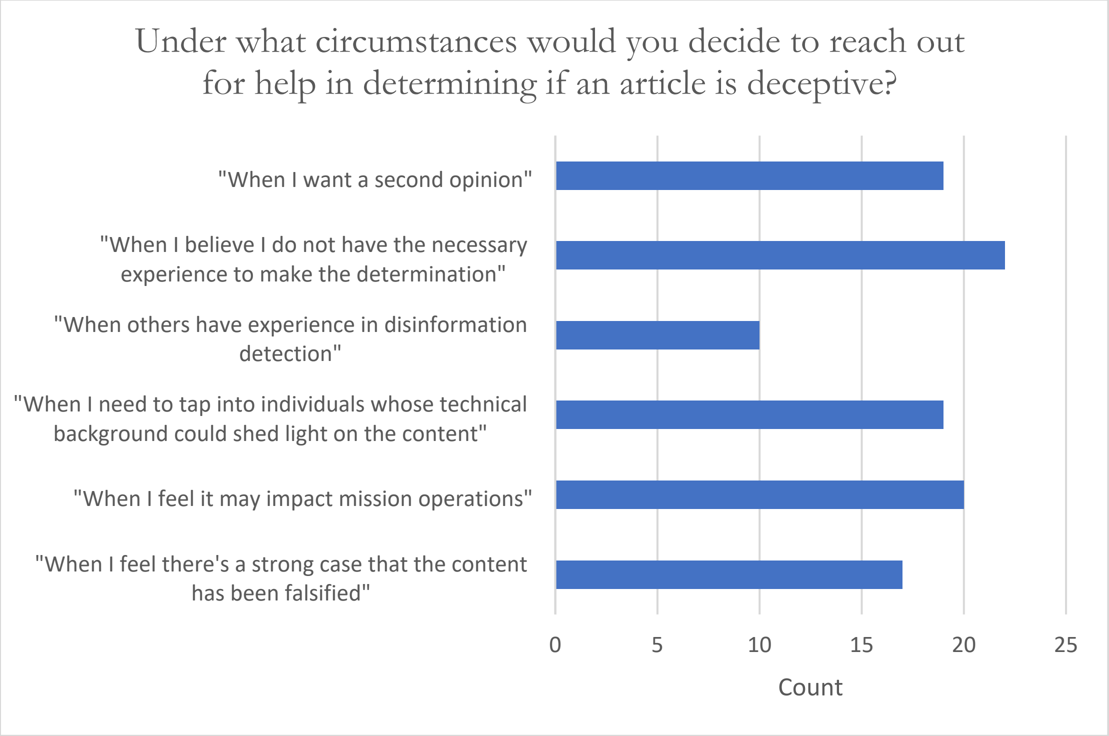


Figure 8. Review Strategy – When to Reach Out for Help?

3.6. Article Linguistic Analysis

A series of planned t-tests were performed for each measure of the 160 measures returned from Receptiviti comparing text in retracted and not retracted articles. None of these tests achieved significance (p's ≥ 0.1), likely due to only having 8 articles in each class. Figure 9 and Figure 10 present a reporting of selected Receptiviti measures of primary interest. For the linguistic cues with stronger predictions based on prior literature (see section 1.3.2), many of these measures did have numerical patterns in the expected direction. For instance, retracted articles had longer sentences on average than non-retracted articles and also featured more jargon (indicated by fewer proportion of the total words were from the LIWC dictionary), which are hallmarks of obfuscation and reading difficulty. Also, many of the measures that are thought to add specificity and concreteness to the text (e.g. prepositions, articles, conjunctions, adjectives) also showed patterns in the predicted direction, with retracted articles showing relatively lower rates of those parts of speech. This finding aligns with previous studies showing that deceptive writing is often more vague and lacks concreteness.

However, not all predictions were met – for instance, words were not longer (more complex) in retracted articles than non-retracted. Either way, these observations reflect only numerical trends. With a larger sample of retracted and non-retracted articles, more definitive conclusions would be able to be reached.

Dictionary Measures

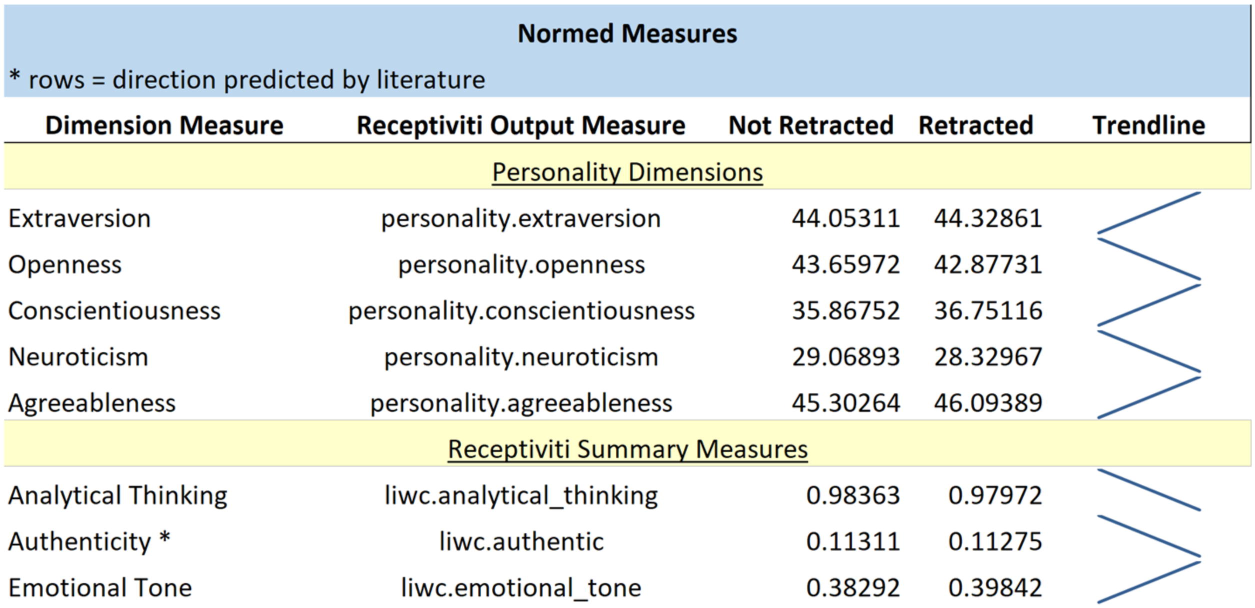
* rows = direction predicted by literature

Dimension Measure	Receptiviti Output Measure	Not Retracted	Retracted	Trendline
Complexity Measures				
Words per sentence *	summary.words_per_sentence	21.49873	25.32581	↗
Big words	liwc.six_plus_words	0.35989	0.34904	↘
Dictionary words *	liwc.dictionary_words	0.65348	0.65183	↘
Parts of Speech Measures				
1st person pronouns				
"I"	liwc.i	0.01945	0.02093	↗
"we" *	liwc.we	0.00094	0.00042	↘
2nd/3rd person pronouns				
"you"	liwc.you	0.00497	0.0025	↘
"she/he"	liwc.she_he	0.00007	0.00004	↘
"they"	liwc.they	0.00009	0.00002	↘
Prepositions *	liwc.prepositions	0.15018	0.14311	↘
Articles *	liwc.articles	0.08997	0.0883	↘
Adverbs *	liwc.adverbs	0.01623	0.01766	↗
Conjunctions *	liwc.conjunctions	0.04815	0.04257	↘
Common adjectives *	liwc.adjectives	0.04061	0.04044	↘
Conceptual Word Measures				
Quantifiers	liwc.quantifiers	0.01976	0.0207	↗
Comparisons *	liwc.comparisons	0.02576	0.02456	↘
Differentiation *	liwc.differentiation	0.01662	0.01354	↘
All-or-none (absolutist)	liwc_extension.absolutist	0.00217	0.00243	↗
Causation *	liwc.causation	0.03252	0.03082	↘
Tentative *	liwc.tentative	0.00868	0.0097	↗
Certainty *	liwc.certainty	0.00728	0.00724	↘
(+) Emotion Words	liwc.positive_emotion_words	0.0138	0.01598	↗
(-) Emotion Words *	liwc.negative_emotion_words	0.0071	0.00895	↗
Abstraction	liwc_extension.abstract	0.32628	0.31742	↘
Concreteness	liwc_extension.concrete	0.16199	0.16668	↗

Figure 9. Linguistic Analysis Results – Dictionary Counted Measures

In looking at the personality dimensions in the "Normed Measures" figure below, it is worth noting that the dimensions represented below are generally used as a means of describing the personality traits of individuals through self-report responses to questions. The Receptiviti platform seeks to provide a measure for how those personality dimensions are revealed through word choices in text.

There were no compelling predictions for how the manifestation of those personality dimensions in text would differ based on if the article was eventually retracted (and thus potentially deceptive) or not. There were no significant differences between retracted and non-retracted papers on the Personality or Receptiviti Summary dimensions. It is interesting to note that the measure for "authenticity" was lower for retracted papers, as the authenticity measure was initially developed by Receptiviti to be a summary measure reflecting authenticity versus deception based on characteristics of deceptive writing in the literature (e.g., Newman et al 2003).



Normed Measures

* rows = direction predicted by literature

Dimension Measure	Receptiviti Output Measure	Not Retracted	Retracted	Trendline
Personality Dimensions				
Extraversion	personality.extraversion	44.05311	44.32861	
Openness	personality.openness	43.65972	42.87731	
Conscientiousness	personality.conscientiousness	35.86752	36.75116	
Neuroticism	personality.neuroticism	29.06893	28.32967	
Agreeableness	personality.agreeableness	45.30264	46.09389	
Receptiviti Summary Measures				
Analytical Thinking	liwc.analytical_thinking	0.98363	0.97972	
Authenticity *	liwc.authentic	0.11311	0.11275	
Emotional Tone	liwc.emotional_tone	0.38292	0.39842	

Figure 10. Linguistic Analysis Results – Normed Measures

3.7. ABM

Of additional interest was if the current Agent-Based Model (ABM) design and parameterization based on the empirical human psychological and behavioral data could effectively replicate the performance observed by our human participants. A more dynamic and complex individual agent structure was used to be more fully representative of those features (both of humans and written content) most impactful for deception and disinformation detection. As such, one of the main scenarios explored by the model sought to simulate the deception detection rates and performance by our participant group. It was demonstrated that the ABM is generally able to emulate the observed human performance well, with both the ABM simulation and our participants achieving roughly a 50% detection rate. While this may seem indicative of performance at chance, this is not an uncommon performance level for humans when faced with detecting deception or lies (Bond & DePaulo 2006). While there are marked differences in the exact distribution and variability of performance scores (i.e., "agents" within the model did not exactly replicate the kind of human data observed), consistency of the overall mean performance indicates some level of success in being able to effectively build an ABM with more complex agents that can replicate human behavior at a general level within the context of disinformation detection. Additional results, analyses, and scenarios are outlined and discussed in SAND2023-08981 (Emery et al. 2023).

3.8. Limitations

3.8.1. Limited data

The power of statistical tests for this work was limited by the number of trials each participant performed (four, one outcome measure of accuracy for each evaluated article), as well as the total number of participants, which was 23 across all domains. The primary barrier accounting for both limitations was the length of time the experiment took. Requesting participants to return for two additional sessions of 2.5 hours each means an individual participant would have spent around six hours of time to complete the study across all three sessions. In an ideal design, each participant would have spent only an hour of experiment time on a behavioral task, and would have completed many more, shorter "trials." This would likely translate to a substantial increase to the number of trials per participant and six times the number of participants as the current sample for the same total aggregate experimental time.

There were, of course, trade-offs with developing the methodology. The long experimental time was necessary for individual participants to "deep dive" their analysis of full articles. An alternative design could have participants read only paper abstracts or curated short excerpts from full articles. This would have enabled more trials, but it would have failed to capture the experience of full article evaluation. Additionally, as most retracted publications were multi-author, a design using article excerpts introduces the potential for a selected excerpt to not contain any contributions from a deliberately deceptive author, instead being a passage written by an innocent colleague (see 4.3.2 below for more). Another alternate design could involve periodically querying whether the article contained potential deception (e.g., submit an evaluation after each major section of an article), but this design has a shortcoming of potentially interfering with typical article reading and introducing undesirable psychological effects.

3.8.2. Multiple author complications

Analyses of documents for deceptive language assumes that authors intend to deceive. However, although all papers were retracted for intentional falsehoods (e.g., data fabrication and manipulation) when multiple authors are involved each individual author's awareness of this fraud is challenging to determine from retraction notices. Thus, it is possible that subsections or entire documents may have been produced by team members who were not party to the deception of their colleagues. For instance, a graduate student author may fabricate data and only produce content for the results section while their naïve colleagues wrote all the supporting introductory and discussion content. Single author papers would avoid this conundrum, but they are also exceptionally rare in the current academic publishing climate. Overall, the issue of uncertainty in text provenance across authors only serves to make it more challenging to obtain results comparing non-deceptive to deceptive articles. Any reported results were obtained despite this challenge and not because of it.

3.8.3. Previous work - identifying linguistic indicators in "paper mill" documents

It might appear that this work does not obtain statistically significant differences across retracted and non-retracted papers based on linguistic cues whereas previous work at Sandia has reportedly been able to. However, we note the following distinctions between this and previous work (in addition to the limitation in 3.8.1 regarding limited numbers of articles):

- Previous motivating work at Sandia was able to correctly classify fraudulent academic papers based on POS and compression based algorithmic classification, not traditional inferential statistics
- The fraudulent papers collected in some of the previous work were thought to all come from the same "paper mill" (Bik 2020) as opposed to disparate and unrelated authors.
- Algorithmic classification in the above case may be using regularities characteristic of the way the specific "paper mill" generated those fraudulent technical reports, rather than deception-related linguistic cues.

This page left blank

4. SUMMARY AND CONCLUSIONS

A comprehensive study was conducted examining how subject matter experts evaluate academic writing for potentially deceptive content, measuring both individual features of the experts and individual features of the documents they examined. With respect to their overall performance, they were at chance, unable to discern which articles had been deceptive (retracted) and which were not. Human frailty at lie detection is common (Bond & DePaulo 2006), and although domain expertise can sometimes be protective in detection of disinformation online (Zrnec et al. 2022), clearly this is not the case for examination of formal technical articles. Additionally, their performance featured a strong truth bias, also consistent with the literature (e.g., Levine et al. 1999), and sensible because most of the time when one reads an academic article, they do not encounter deliberate deception or fraud (Parkinson & Wykes 2023). Of critical interest was not just overall performance, but whether any facets of personality or individual differences in experts would be predictive of better or worse performance, as well as if there were any predictive indicators of fraud in the language used in the retracted documents themselves. Findings related to those topics follow.

With respect to individual differences, a battery of assessments and surveys were conducted (2.3.3) based on prior literature in lie detection and disinformation discernment. Only a few measures were predictive of behavior: Need for Closure (NFC, Roets & Van Hiel 2011), the Comprehensive Intellectual Humility "lack of intellectual overconfidence" subscale (Krumrei-Mancuso & Rouse 2016), and Shipley Abstract (Kaya et al. 2012; Shipley et al. 2009). Need for closure, which measures an individual's comfort level with uncertainty and ambiguity (higher scores reflecting a need for clear resolutions), was predictive in the expected direction, with those having lower scores performing better at discerning which articles contained deception. Similarly, the "lack of intellectual overconfidence" subscale is intended to reflect an individual's ability to recognize that they might not know everything about a topic. As expected, people scoring higher on this subscale, who demonstrated appropriate humility for their own knowledge relative to the potential knowledge of others, also performed better on the task. Notably, performance on these two assessments were negatively correlated (2.4.2.1), such that those with lower NFC scores had higher scores on the "lack of intellectual overconfidence" subscale.

Together, these findings suggest that a sense of general comfort with uncertainty might be important for the ability to discern that an article contained deceptive content. However, other cognitive factors that might also be considered important for general, and thorough, critical thinking were not predictive (e.g., the Cognitive Reflection Test and the Big 5 personality measure of conscientiousness). Finally, that higher scores in abstract intelligence (e.g., pattern completion) were predictive of *lower* performance on the task suggests that appropriate caution should be taken to potentially over-interpret these outcomes. This is a relatively small sample of participants (N = 23) who each only evaluated four documents. Future work at a larger scale will be necessary to determine which of these effects is robust.

Finally, one question that remains is, is there even a signal in the writing that would be reliability indicative of deception for readers to be able to discern? The task of deception detection inherently assumes that deception would have indicators that a reader could flag and utilize in their assessment of article validity. While this may be true in machine learning with large sets of data and near-infinite memory capacity to track patterns, the typical scale of documents and information that humans can thoroughly read and process is much smaller. For instance, in this study it took each participant several hours to deep-dive only four articles. The findings from the linguistic cue analysis (3.6), where no significant differences were obtained across dozens of potential indicators with this limited set of articles (8 retracted, 8 non-retracted), suggests that there were no stand-out cues that an individual could reliability monitor for deception in the article, even if they were told what to look for. It is one thing for machines to be able to pick up on these subtle statistical regularities, but it is quite possible that few or none of these linguistic differences can rise to the level of conscious awareness in humans. That is, the cues of deception that are identifiable in academic writing (e.g., Markowitz & Hancock 2014, 2016) generally depend on a larger set of materials and processing capacity than what a human could reasonably be expected to process for reliable differences to be obtained. No human is going to manually count the number of adverbs in an article, nor should they – this is a task better suited to a machine.

One potential suggestion for the identification of academic fraud in published articles is to combine the strength of machine computational power with the human ability to evaluate contextual information (Does this person publish too frequently and get too many "surprising" outcomes? Do the claims the author is making follow from the data? Is this methodology consistent?). In this manner, machine down-sampling could red-flag potential fraud, and human experts would subject the identified articles – and the authors who generated those articles – to further scrutiny.

In conclusion, this study provides valuable insights concerning both experts' abilities to detect potentially deceptive content in formal technical writing, as well as the psychological characteristics of those experts. These findings provide valuable context and data that can be used to both inform mission problems and motivate future deception and disinformation research.

REFERENCES

1. Ajina, A., Laouiti, M., & Msolli, B. (2016). Guiding through the fog: does annual report readability reveal earnings management?. Research in International Business and Finance, 38, 509-516.
2. Bik, E. (2020, July 19). The Stock Photo Paper Mill. Science Integrity Digest. https://scienceintegritydigest.com/2020/07/05/the-stock-photo-paper-mill/
3. Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and social psychology Review, 10(3), 214-234.
4. Bond Jr, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: accuracy and bias. Psychological bulletin, 134(4), 477.
5. Bowes, S. M., & Tasimi, A. (2022). Clarifying the relations between intellectual humility and pseudoscience beliefs, conspiratorial ideation, and susceptibility to fake news. Journal of Research in Personality, 98, 104220.
6. Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin.
7. Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., ... & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American journal of public health, 108(10), 1378-1384.
8. Brumfiel, G., & McMinn, S. (2018, December 19). Open Scientific Collaboration May Be Helping North Korea Cheat Nuclear Sanctions. NPR. https://www.npr.org/2018/12/19/675390104/open-scientific-collaboration-may-be-helping-north-korea-cheat-nuclear-sanctions
9. Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. *PLoS One*, *16*(6), e0253717.
10. Buchanan, T. (2020). Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. Plos one, 15(10), e0239666.
11. Buchanan, T. (2021). Trust, personality, and belief as determinants of the organic reach of political disinformation on social media. The Social Science Journal, 1-12.
12. Buchanan, T., & Benson, V. (2019). Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of "fake news"?. Social media+ society, 5(4), 2056305119888654.
13. Burgoon, J. K., Blair, J. P., & Strom, R. E. (2008). Cognitive biases and nonverbal cue availability in detecting deception. Human Communication Research, 34(4), 572-599.
14. Calvillo, D. P., Garcia, R. J., Bertrand, K., & Mayers, T. A. (2021). Personality factors and self-reported political news consumption predict susceptibility to political fake news. Personality and individual differences, 174, 110666.
15. Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. Annu. Rev. Psychol., 56, 453-484.
16. Checkbox Technology, Inc. Checkbox survey solutions, 2023.
17. COPE. (2020). Potential "paper mills" and what to do about them – A publisher's perspective. https://publicationethics.org/publishers-perspective-paper-mills
18. Costa, P. T., & McCrae, R. R. (1999). A five-factor theory of personality. The five-factor model of personality: Theoretical perspectives, 2, 51-87.

19. Curtis, D. A. (2021). Deception detection and emotion recognition: Investigating FACE software. Psychotherapy Research, 31(6), 802-816.
20. De Keersmaecker, J. & Roets, A. (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. Intelligence, 65, 107-110.
21. de Souza, J. A. S., Rissatti, J. C., Rover, S., & Borba, J. A. (2019). The linguistic complexities of narrative accounting disclosure on financial statements: An analysis based on readability characteristics. Research in International Business and Finance, 48, 59-74.
22. Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. Personality and Social Psychology Review, 14(2), 238-257.
23. DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. Personality and Social Psychology Review, 1(4), 346-357.
24. DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-Job Experience and Skill at Detecting Deception 1. Journal of Applied Social Psychology, 16(3), 249-267.
25. Duffy, J. (2006). Agent-based models and human subject experiments. Handbook of computational economics, 2, 949-1011.
26. Emery, B., S. Verzi, D. Dickson, & T. Gunda. 2023. "Discerning Deception: An Empirically-Driven Agent-Based Model of Expert Evaluation of Scientific Content." Sandia National Laboratories, SAND2023-08981.
27. Evans, A., Sleegers, W., & Mlakar, Ž. (2020). Individual differences in receptivity to scientific bullshit. Judgment and Decision Making, 15(3), 401-412.
28. Fallis, D. (2014). The varieties of disinformation. The philosophy of information quality, 135-161.
29. Fallis, D. (2015). What is disinformation?. Library trends, 63(3), 401-426.
30. Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences, 109(42), 17028-17033.
31. Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19, 25–42.
32. Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. Journal of Verbal Learning and Verbal Behavior, 21(2), 207-219.
33. Hartwig, M., & Bond Jr, C. F. (2014). Lie detection from multiple cues: A meta-analysis. Applied Cognitive Psychology, 28(5), 661-676.
34. Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. Journal of verbal learning and verbal behavior, 16(1), 107-112.
35. Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Personality and social psychology Review, 19(4), 307-342.
36. Humpherys, S. L. (2009). Discriminating fraudulent financial statements by identifying linguistic hedging. AMCIS 2009 Proceedings, 400.
37. Jalbert, M., Newman, E., & Schwarz, N. (2020). Only half of what I'll tell you is true: Expecting to encounter falsehoods reduces illusory truth. Journal of Applied Research in Memory and Cognition, 9(4), 602-613.

38. Jankowicz, N. (2020). How to lose the information war: Russia, fake news, and the future of conflict. Bloomsbury Publishing.
39. JASP Team (2023). JASP (Version 0.17.3)[Computer software].
40. Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological bulletin, 114(1), 3.
41. Kaligotla, C., Yücesan, E., & Chick, S. E. (2022). Diffusion of competing rumours on social media. Journal of Simulation, 16(3), 230-250.
42. Kaminska, I. (2017). A lesson in fake news from the info-wars of ancient Rome. Financial Times, 17.
43. Kaya, F., Delen, E., & Bulut, O. (2012). Test Review: Shipley-2 Manual. Journal of Psychoeducational Assessment, 30(6), 593–597.
44. Knapp, M. L., & Comaden, M. E. (1979). Telling it like it isn't: A review of theory and research on deceptive communications. Human Communication Research, 5(3), 270-285.
45. Knapp, Mark L., Roderick P. Hart, & Harry S. Dennis. "An exploration of deception as a communication construct." Human communication research 1, no. 1 (1974): 15-29.
46. Koetke, J., Schumann, K., & Porter, T. (2022). Intellectual humility predicts scrutiny of COVID-19 misinformation. Social Psychological and Personality Science, 13(1), 277-284.
47. Krumrei-Mancuso, E. J., & Rouse, S. V. (2016). The development and validation of the comprehensive intellectual humility scale. Journal of Personality Assessment, 98(2), 209-221.Big 5
48. Levine, T. R. (2015). New and improved accuracy findings in deception detection research. Current Opinion in Psychology, 6, 1-5.
49. Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". Communications Monographs, 66(2), 125-144.
50. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological science in the public interest, 13(3), 106-131.
51. Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348-384.
52. Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics, 45(2-3), 221-247.
53. Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. Journal of accounting and Economics, 63(1), 1-25.
54. Marchlewska, M., Cichocka, A., & Kossowska, M. (2018). Addicted to answers: Need for cognitive closure and the endorsement of conspiracy beliefs. European journal of social psychology, 48(2), 109-117.
55. Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. PloS one, 9(8), e105937.
56. Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. Journal of Language and Social Psychology, 35(4), 435-445.
57. Markowitz, D. M., Kouchaki, M., Hancock, J. T., & Gino, F. (2021). The deception spiral: Corporate obfuscation leads to perceptions of immorality and cheating behavior. Journal of Language and Social Psychology, 40(2), 277-296.

58. Markowitz, D. M., Powell, J. H., & Hancock, J. T. (2014, June). The writing style of predatory publishers. In 2014 ASEE Annual Conference & Exposition (pp. 24-1259).
59. McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. American psychologist, 52(5), 509.
60. Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. Law and human behavior, 26, 469-480.
61. Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. Nature communications, 12(1), 921.
62. Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: a review of research in social media. International journal of data science and analytics, 13(4), 271-285.
63. Nemr, C., & Gangware, W. (2019). Weapons of mass distraction: Foreign state-sponsored disinformation in the digital age. Park Advisors.
64. Newman, E. J., Jalbert, M. C., Schwarz, N., & Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of need for cognition. Consciousness and Cognition, 78, 102866.
65. Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, 29(5), 665-675.
66. Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: a randomized trial. Pediatrics, 133(4), e835-e842.
67. Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. Vaccine, 33(3), 459-464.
68. Olcott, E., Smith, A., & Cookson, C. (2023, March 28). China's fake science industry: how 'paper mills' threaten progress. Retrieved from https://www.ft.com/content/32440f74-7804-4637-a662-6cdc8f3fba86
69. Oransky, I. & Marcus, A. Retraction watch database, 2023.
70. Parkinson, A., & Wykes, T. (2023). The anxiety of the lone editor: fraud, paper mills and the protection of the scientific record. Journal of Mental Health, 32(5), 865-868.
71. Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. Annual review of psychology, 54(1), 547-577.
72. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science, 31(7), 770-780.
73. Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition, 188, 39-50.
74. Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of personality, 88(2), 185-200.
75. Posetti, J., & Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. International Center for Journalists, 7(2018), 2018-07.
76. Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality, 41(1), 203-212.
77. Receptiviti Inc. (2022). *receptiviti: Text Analysis Through the Receptiviti API.* https://receptiviti.github.io/receptiviti-r/.

78. The Retraction Watch Database [Internet]. New York: The Center for Scientific Integrity. 2018. ISSN: 2692-465X. [Cited August 2022]. Available from: http://retractiondatabase.org/.
79. Roberts, B. W., & Yoon, H. J. (2022). Personality psychology. Annual review of psychology, 73, 489-516.
80. Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. Personality and individual differences, 50(1), 90-94.
81. Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). Shipley-2. Los Angeles, CA: Western Psychological Services.
82. Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), e1385.
83. Sindermann, C., Cooper, A., & Montag, C. (2020). A short review on susceptibility to falling for fake political news. Current Opinion in Psychology, 36, 44-48.
84. Sindermann, C., Schmitt, H. S., Rozgonjuk, D., Elhai, J. D., & Montag, C. (2021). The evaluation of fake and true news: on the role of intelligence, personality, interpersonal trust, ideological attitudes, and news consumption. Heliyon, 7(3).
85. Smith, N. (2001). “Reading Between the Lines: An Evaluation of the Scientific Content Analysis Technique (SCAN),” in C. F. Willis (ed.), Policing and Reducing Crime Unit: Police Research Series, London: Crown.
86. Tausczik, Yia R., and James W. Pennebaker. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." Journal of Language and Social Psychology 29, no. 1 (2010): 24-54.
87. Wall, H. J., Campbell, C. C., Kaye, L. K., Levy, A., & Bhullar, N. (2019). Personality profiles and persuasion: An exploratory study investigating the role of the Big-5, Type D personality and the Dark Triad on susceptibility to persuasion. Personality and Individual Differences, 139, 69-76.
88. Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. Journal of personality and social psychology, 67(6), 1049.
89. Wolverton, C. and Stevens, D. (2020), "The impact of personality in recognizing disinformation", Online Information Review, 44(1), pp. 181-191. https://doi.org/10.1108/OIR-04-2019-0115
90. Zloteanu, M., Bull, P., Krumhuber, E. G., & Richardson, D. C. (2021). Veracity judgement, not accuracy: Reconsidering the role of facial expressions, empathy, and emotion recognition training on deception detection. Quarterly Journal of Experimental Psychology, 74(5), 910-927.
91. Zrnec, A., Poženel, M., & Lavbič, D. (2022). Users’ ability to perceive misinformation: An information quality assessment approach. Information Processing & Management, 59(1), 102739.

This page left blank

APPENDIX A. BEHAVIORAL STUDY ARTICLE REFERENCES

Retracted

- Mori, N., Wada, A., Hirayama, T., Parks, T. P., Stratowa, C., & Yamamoto, N. (2000). Activation of intercellular adhesion molecule 1 expression by Helicobacter pylori is regulated by NF-ϰB in gastric epithelial cancer cells. Infection and immunity, 68(4), 1806-1814.
- Khan, F., Kumari, M., & Cameotra, S. S. (2013). Biodegradation of the allelopathic chemical m-tyrosine by Bacillus aquimaris SSC5 involves the homogentisate central pathway. Plos one, 8(10), e75928.
- Guerra, D. L., Viana, R. R., & Airoldi, C. (2009). RETRACTED: Designed pendant chain covalently bonded to analogue of heulandite for removal of divalent toxic metals from aqueous solution: Thermodynamic and equilibrium study.
- Casciato, M. J., Levitin, G., Hess, D. W., & Grover, M. A. (2012). Synthesis of optically active ZnS–carbon nanotube nanocomposites in supercritical carbon dioxide via a single source diethyldithiocarbamate precursor. Industrial & engineering chemistry research, 51(36), 11710-11716.
- Saha, P., Mukherjee, D., Singh, P. K., Ahmadian, A., Ferrara, M., & Sarkar, R. (2021). Retracted article: Graphcovidnet: A graph neural network based model for detecting COVID-19 from ct scans and x-rays of chest. Scientific reports, 11(1), 8304.
- Chen, H., Song, M., Zhao, J., Dai, Y., & Li, T. (2019, June). Retracted on January 26, 2021: 3D-based video recognition acceleration by leveraging temporal locality. In Proceedings of the 46th International Symposium on Computer Architecture (pp. 79-90).
- Ghaffari, M., Zhou, Y., Xu, H., Lin, M., Kim, T. Y., Ruoff, R. S., & Zhang, Q. M. (2013). High‐Volumetric Performance Aligned Nano‐Porous Microwave Exfoliated Graphite Oxide‐based Electrochemical Capacitors. Advanced Materials, 25(35), 4879-4885.
- Mahato, P., Yanai, N., Sindoro, M., Granick, S., & Kimizuka, N. (2016). Preorganized chromophores facilitate triplet energy migration, annihilation and upconverted singlet energy collection. Journal of the American Chemical Society, 138(20), 6541-6549.

Not Retracted

- Innocenti, M., Thoreson, A. C., Ferrero, R. L., Stromberg, E., Bolin, I., Eriksson, L., ... & Quiding-Jarbrink, M. (2002). Helicobacter pylori-induced activation of human endothelial cells. Infection and immunity, 70(8), 4581-4590.
- Bertin, C., Weston, L. A., Huang, T., Jander, G., Owens, T., Meinwald, J., & Schroeder, F. C. (2007). Grass roots chemistry: meta-tyrosine, an herbicidal nonprotein amino acid. Proceedings of the National Academy of Sciences, 104(43), 16964-16969.
- Dey, R. K., & Airoldi, C. (2008). Designed pendant chain covalently bonded to silica gel for cation removal. Journal of hazardous materials, 156(1-3), 95-101.
- Wang, Y., Liu, Z., Han, B., Zhang, J., Jiang, T., Wu, W., ... & Hang, Y. (2003). Synthesis of polypropylene/ZnS composite using the template prepared by supercritical CO2. Chemical physics letters, 381(3-4), 271-277.

- Fouladi, S., Ebadi, M. J., Safaei, A. A., Bajuri, M. Y., & Ahmadian, A. (2021). Efficient deep neural networks for classification of COVID-19 based on CT images: Virtualization via software defined radio. Computer communications, 176, 234-248.
- Zhou, C., Liu, M., Qiu, S., He, Y., & Jiao, H. (2021, December). An Energy-Efficient Low-Latency 3D-CNN Accelerator Leveraging Temporal Locality, Full Zero-Skipping, and Hierarchical Load Balance. In 2021 58th ACM/IEEE Design Automation Conference (DAC) (pp. 241-246). IEEE.
- Hosoyamada, M., Yanai, N., Ogawa, T., & Kimizuka, N. (2016). Molecularly dispersed donors in acceptor molecular crystals for photon upconversion under low excitation intensity. Chemistry–A European Journal, 22(6), 2060-2067.
- Murali, S., Quarles, N., Zhang, L. L., Potts, J. R., Tan, Z., Lu, Y., ... & Ruoff, R. S. (2013). Volumetric capacitance of compressed activated microwave-expanded graphite oxide (a-MEGO) electrodes. Nano Energy, 2(5), 764-768.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Matthew Windsor	5571	mwindso@sandia.gov
Danielle Dickson	5572	dsdicks@sandia.gov
Thushara Gunda	8932	tgunda@sandia.gov
Curtis Johnson	5552	cjohnso@sandia.gov
Nicole Murchison	5523	nmurchi@sandia.gov
Jason Morris	5571	jmorris@sandia.gov
Susan Adams	5572	smsteve@sandia.gov
Danielle Sanchez	5571	dnsanc@sandia.gov
Aaron Jones	5572	ajones3@sandia.gov
Brad Robert	5572	bmrober@sandia.gov
Emily Kemp	5573	ekemp@sandia.gov
Alisa Rogers	5572	anroger@sandia.gov
Shanle Longmire-Monford	5571	selongm@sandia.gov
Kyle Selasky	5571	kcselas@sandia.gov
Kurtis Shuler	5573	kwshule@sandia.gov
Lawrence Allen	9734	lcallen@sandia.gov
Marie Tuft	5574	mtuft@sandia.gov
Technical Library	1911	sanddocs@sandia.gov

This page left blank

This page left blank

Sandia National Laboratories

Sandia National Laboratories
is a multimission laboratory
managed and operated by
National Technology &
Engineering Solutions of
Sandia LLC, a wholly owned
subsidiary of Honeywell
International Inc. for the U.S.
Department of Energy's
National Nuclear Security
Administration under contract
DE-NA0003525.