International Journal of Cybersecurity Intelligence & Cybercrime

Volume 4 | Issue 1 Article 2

4-5-2021

Assessing the Credibility of Cyber Adversaries

Recommended Citation

Wells, J. A., LaFon, D. S., & Gratian, M. (2021). Assessing the Credibility of Cyber Adversaries, International Journal of Cybersecurity Intelligence & Cybercrime, 4(1), 3-24. https://www.doi.org/10.52306/04010221FHTE2115

Wells, J., LaFon, D., & Gratian, M. (2021). *International Journal of Cybersecurity Intelligence and Cybercrime, 4*(1), 3-24.

Assessing the Credibility of Cyber Adversaries

Jenny A. Wells*, Australian Signals Directorate, Australia
Dana S. LaFon, National Security Agency, U.S.A.
Margaret Gratian, National Security Agency, U.S.A.

Keywords; Credibility, trust, deception, online credibility

Abstract:
Online communications are ever increasing, and we are constantly faced with the challenge of whether online information is credible or not. Being able to assess the credibility of others was once the work solely of intelligence agencies. In the current times of disinformation and misinformation, understanding what we are reading and to who we are paying attention to is essential for us to make considered, informed, and accurate decisions, and it has become everyone's business. This paper employs a literature review to examine the empirical evidence across online credibility, trust, deception, and fraud detection in an effort to consolidate this information to understand adversary online credibility – how do we know with whom we are conversing is who they say they are? Based on this review, we propose a model that includes examining information as well as user and interaction characteristics to best inform an assessment of online credibility. Limitations and future opportunities are highlighted.

Introduction

In October 2020, 59% of the world's population were active internet users. In North America and Northern Europe, 95% of the population use the internet (Clement, 2020). With the ever-increasing use of cyber activity and user-generated content, people regularly face the question "Is this real?" The notion of "fake news," popularized by the 2016 U.S. presidential campaign, has become an overarching concept for how facts and information are gathered from online sources (Egelhofer & Lecheler, 2019) and is used to misrepresent false material and false sources as credible. In addition to false misrepresentations, the rising use of phishing and other cybercrimes aimed at the general public, organizations, and governments have brought the concept of online credibility into current public discourse.

The escalating use of technology to conduct malicious activities by foreign intelligence services, state-sponsored actors, and criminal actors raises grave concerns about online credibility. One of the most recent examples of mass malicious cyber activities targeting the public occurred during the 2016 U.S. election. The report of the U.S. investigation cited: "The Internet Research Agency [IRA] carried out the earliest Russian interference operations identified by the investigation…..The IRA later carried used social media accounts and interest groups to sow discord in the U.S. political system through what it termed 'information warfare". The campaign evolved from a generalized program designed in 2014 and 2015 to undermine the U.S. electoral system, to a targeted operation that by early 2016 favored candidate Trump and disparaged candidate Clinton. The IRA's operation also included the purchase of political advertisements on social media in the names of U.S. persons and entities, as well as the staging of political rallies inside the United States" (Mueller, Helderman, & Zapotosky, 2019, p. 4).

*Corresponding author
Jenny Wells, Australian Signals Directorate, Russell Offices, Russell, ACT, 2600, Australia
Email: jennyawells@gmail.com

Using tried and tested techniques of developing credible personas, the state-sponsored campaign leveraged the credibility of real Americans through the liking and sharing of Russian originated persona posts and online activities, which sought to sow discord, increase uncertainty, and magnify social tensions to influence voting behaviour (Jamieson, 2018). The reach of this campaign was extensive. For example, in 2018, Twitter notified 1.4 million Twitter users who had directly engaged or actively followed the now-identified 3,841 IRA-linked accounts (Twitter Public Policy, 2018).

However, political advantage is not the only target of cyber adversaries. While other targets may include government agencies and businesses, everyday internet citizens are frequently the target. The Australian Cyber Security Centre (ACSC; 2020) identified that the most common cybercrime perpetrated against individuals is fraud by deception followed by identity-theft. The ACSC noted that cyber adversaries also commonly conduct phishing and spear phishing campaigns, along with corporate email compromise. Purplesec (2020) reported that corporate email compromises cost organizations $676 million in 2017 and note that 98% of all cyber-attacks rely on social engineering. As such, this is no longer an intelligence service problem; the general population must constantly question the authenticity of online identities and information from potential adversaries, which is any online source intended to introduce false, misleading, or inaccurate information for nefarious or malicious purposes. This paper aims to apply theoretical and empirical advances in understanding online credibility of malicious cyber actors through a model of online adversary credibility assessment.

Literature Review and Methodology

The literature review process entailed a thorough search of relevant scientific databases to include, but not limited to, the EBSCO database. A vast array of pertinent search terms was used to identify research focusing on assessing credibility, trust, and deception, among other relevant topics. Search terms included (with synonyms and closely related words); credibility, credibility assessment, trustworthiness, believability, digital trust, online behavior, online communication, cyber adversaries, trust, cybercrime, online deception, and fraud detection. Initially articles between 2005 and 2020 were reviewed; however, some additional studies that were relevant to the current article were published outside of the initial date range. Further studies were identified by examining the reference lists of all included articles and searching relevant websites.

Information Processing

As humans, we are influenced by our emotions and thoughts and the way we process information is not always considered, thoughtful, logical, or rational. The dual process theory of information processing stipulates that there are two main ways that we process information - one is an automatic process, and one is a slow, thoughtful process (Kahneman & Frederick, 2005). The former is known as System 1 thinking - it is fast thinking that is effortless and often emotion and heuristic driven. The latter is known as System 2 thinking, which requires cognitive resources, uses abstract thinking, and is both critical and logical. Both System 1 and System 2 are useful in different situations (e.g., driving a car versus calculating a complex mathematical problem).

Understanding the way that humans process information helps us understand how online credibility assessments are made. Research indicates that System 1 thinking dominates our information processing,

whereby we prefer this quick thinking to effortful consideration (Sundar, Knobloch-Westerwick, & Hastall, 2007). Malicious cyber actors rely on the vulnerabilities of System 1 thinking to manipulate people to the desired behavior. Research has focused on identifying factors that can increase the susceptibility and vulnerability to online influence and System 1 thinking, which includes familiarity (Begg, Anas, & Farinacci, 1992), emotional triggers (Langenderfer & Shimp, 2001), salience of information (Igartua & Cheng, 2009), perceived credibility (Pornpitakpan, 2004), and propensity to trust (Bond & DePaulo, 2006).

Engaging System 1 processing is both time and energy efficient but increases the risk of poor decision-making, biased thinking, and impaired judgement (Tversky & Kahneman, 1974; Vishwanath, Harrison, & Ng, 2018). In addition, when dealing with a constant stream of online information, relying on System 1 thinking can increase the likelihood of deception being efficacious and unnoticed on social media (Vishwanath, 2015). While System 2 thinking can ensure that we make slow, deliberate decisions, it is not time or energy efficient. However, identifying when to engage System 2 processing and applying scrutiny and critical thinking to online information can increase the likelihood of identifying deception and manipulation by malicious cyber actors.

How humans process information is fundamental to assessing online credibility. Research has shown that there are roles for both System 1 and 2 thinking in assessing credibility (Metzger & Flanagin, 2015; Sundar, 2008). The following models of online credibility incorporate both the use of Systems 1 and 2 in evaluating online credibility. Our ability to engage System 2 thinking in online credibility assessment is limited by our experience and knowledge about contraindicators for credibility. This paper explores the scientific research on online credibility to increase our knowledge of System 2 thinking engagement and to propose an empirically based model of assessing online adversary credibility.

Credibility. In early work exploring the quality and credibility of information, Taylor (1986) suggested that people form judgments about information through assigning value to some and not to others. This then helps people make decisions about what information to use, share, and inform our actions. Since this time, researchers have examined the underlying factors of this 'judgment'.

In credibility literature, two major dimensions have been found to be related to credibility – trustworthiness and expertise (Fogg & Tseng, 1999; Metzger, 2007). Trustworthiness is being perceived as truthful and honest, and information is considered trustworthy when it appears to be reliable, unbiased, and fair (Hilligoss & Rieh, 2008). Expertise is the perception of one's knowledge, skill, and experience, which is linked to user assessments of validity and accuracy of information. As such, credibility is a subjective assessment of the quality of being trusted and believed.

The role of trust (as opposed to trustworthiness) has also been identified as related to credibility (Hovland, Janis, & Kelley, 1953). Trust refers to a set of beliefs, characteristics, and behaviors associated with the acceptance of risk, vulnerability, interdependence, expectations, insecurity, and action (Talboom & Pierson, 2013), while credibility refers to a perceived quality of a source, which may or may not result in trust (Rieh & Danielson, 2007).

Theoretical Frameworks for Understanding Online Credibility

Theoretical frameworks attempt to explain how people process online information to help them reach

a credibility assessment, recognising that the majority of people cannot attend to and process all the online information (Lang, 2000). The following scientific literature review highlights the most relevant theoretical models of online credibility. These are summarized in Table 1.

Table 1. Summary of Existing Theoretical Frameworks for Understanding Online Credibility

Theoretical Model	Author	Credibility Features
Prominence-Interpretation Theory	Fogg (2003)	• User involvement • Online content • Task of user • User experience • Individual differences
MAIN Model	Sundar (2008)	• Medium of delivery • Source of information • Interaction/activity • Ease of use
Dual Processing Models	Wathen & Burkell (2002)	• Surface level information • Message content
	Metzger (2007)	• Motivation • Ability • User perceptions
Unifying Framework	Hilligoss & Rieh (2008)	• Truthfulness, believability, trustworthiness, objectivity, and reliability • Media, source, endorsement, and aesthetic-based heuristics • Interaction cues
Aggregated Trustworthiness Model	Jessen & Jorgensen (2012)	• Social dynamics
The 3-S Model	Lucassen & Schraagen (2011)	• Information features • Source features • User expertise and experience

Prominence-interpretation theory. This early theory considers two factors in the process of credibility assessment; a person notices something (prominence) and they form a judgment (interpretation).

Prominence-interpretation theory (Fogg, 2003). This early theory considers two factors in the process of credibility assessment; a person notices something (prominence) and they form a judgment (interpretation). Fogg (2003) identified five factors that affect prominence, which includes the involvement of the user (e.g., motivation and ability to process), online content (e.g., type of information, media source), task of user (e.g., seeking amusement, seeking education, making a transaction), user experience (e.g., familiarity with subject matter), and individual differences (e.g., need for cognition, education, learning style, personality). Interpretation occurs once a person notices an online communication and then interprets and assigns value to the message to make a credibility assessment. According to Fogg (2003), interpretation is affected by user assumptions (e.g., culture, beliefs, experiences), skills, and knowledge of user (e.g., level of competency in subject matter), context (e.g., environment, norms, expectations), and user goals (e.g., reasons for online engagement). This process is repetitive for online users, which may result in new cues being noticed and processed.

While this model is useful, it does not encapsulate all individual, technical, and social factors that influence interpretation. Many of these factors, such as the development of various social media and technologies, have evolved since this theory was developed. For example, this model does not consider the influence of other users, the prominence of social media as a source of news, and the development of mobile internet access.

MAIN model (Sundar, 2008). This model focuses on the role of cognitive heuristics in credibility assessments and is primarily concerned with the technological aspects of digital media (Sundar, 2008). Recognising that source, message, and medium are important, the MAIN model looks at the structure of online information to inform credibility judgments. Extensive research has informed this approach, which emphasizes four "affordances" (i.e. Modality, Agency, Interactivity, and Navigability) in digital media that serve to influence credibility assessments through heuristic processing. The author defines affordances as capabilities "that can shape the nature of content in a given medium" (Sundar, 2008, p. 75). Modality relates to the medium of delivery (e.g., credible website appearance, matches expectations with the real world), while Agency relates to the source of the information (e.g., is the source a reputable news company or user-generated content?). Interactivity is defined as interaction and activity. The more interaction and activity people engage in, the more likely they will perceive the information to be credible (e.g., a reviewer on Trip Advisor will perceive the reviews as more credible than someone who has never posted a review). Finally, Navigability is the ease of use and intuitiveness of interface features. Good navigability can trigger heuristic cues of credibility (e.g., provision of hyperlinks, use of navigational aids).

The MAIN model provides an extensive explanation of how digital media can trigger heuristic processing, and how this influences the credibility assessments of such media. However, it focuses on the structural and technical features, rather than the content. It also does not offer explanations for how people engage in analytical thinking processes in assessing credibility.

Dual processing models (Metzger, 2007; Wathen & Burkell, 2002). Dual processing models of credibility assessment focus on how people use credibility indicators in information processing and decision making. Early work recognised the importance of the interaction between source, receiver, and message in a

a credibility assessment (Wathen & Burkell, 2002). They proposed a dual process whereby surface and message factors are assessed to provide an overall credibility evaluation, shown in Table 2.

Fogg (2003) identified five factors that affect prominence, which includes the involvement of the user (e.g., motivation and ability to process), online content (e.g., type of information, media source), task of user (e.g., seeking amusement, seeking education, making a transaction), user experience (e.g., familiarity with subject matter), and individual differences (e.g., need for cognition, education, learning style, personality). Interpretation occurs once a person notices an online communication and then interprets and assigns value to the message to make a credibility assessment. According to Fogg (2003), interpretation is affected by user assumptions (e.g., culture, beliefs, experiences), skills, and knowledge of user (e.g., level of competency in subject matter), context (e.g., environment, norms, expectations), and user goals (e.g., reasons for online engagement). This process is repetitive for online users, which may result in new cues being noticed and processed.

While this model is useful, it does not encapsulate all individual, technical, and social factors that influence interpretation. Many of these factors, such as the development of various social media and technologies, have evolved since this theory was developed. For example, this model does not consider the influence of other users, the prominence of social media as a source of news, and the development of mobile internet access.

MAIN model. This model focuses on the role of cognitive heuristics in credibility assessments and is primarily concerned with the technological aspects of digital media (Sundar, 2008). Recognising that source, message, and medium are important, the MAIN model looks at the structure of online information to inform credibility judgments. Extensive research has informed this approach, which emphasizes four "affordances" (i.e. Modality, Agency, Interactivity, and Navigability) in digital media that serve to influence credibility assessments through heuristic processing. The author defines affordances as capabilities "that can shape the nature of content in a given medium" (Sundar, 2008, p. 75). Modality relates to the medium of delivery (e.g., credible website appearance, matches expectations with the real world), while Agency relates to the source of the information (e.g., is the source a reputable news company or user-generated content?). Interactivity is defined as interaction and activity. The more interaction and activity people engage in, the more likely they will perceive the information to be credible (e.g., a reviewer on Trip Advisor will perceive the reviews as more credible than someone who has never posted a review). Finally, Navigability is the ease of use and intuitiveness of interface features. Good navigability can trigger heuristic cues of credibility (e.g., provision of hyperlinks, use of navigational aids).

The MAIN model provides an extensive explanation of how digital media can trigger heuristic processing, and how this influences the credibility assessments of such media. However, it focuses on the structural and technical features, rather than the content. It also does not offer explanations for how people engage in analytical thinking processes in assessing credibility.

Dual processing models. Dual processing models of credibility assessment (Metzger, 2007; Wathen & Burkell, 2002) focus on how people use credibility indicators in information processing and decision making. Early work recognised the importance of the interaction between source, receiver, and message in a credibility assessment (Wathen & Burkell, 2002). They proposed a dual process whereby surface and message factors are assessed to provide an overall credibility evaluation, shown in Table 2.

According to this model, users first assess surface credibility, whereby the user considers how the online source looks and feels before moving on to assess the source and message credibility. In the final step of this model, the user synthesizes this information with their own previous knowledge to produce an overall credibility assessment.

Table 2. Wathen & Burkell's factors of online credibility (2002)

Surface credibility	Message credibility
Appearance (colors, font, attention to detail)	Source (expertise, trustworthiness, credentials)
Interface design (navigability, interactivity, speed)	Message (content, relevance, currency, accuracy, tailoring)
Organization (layers, ease of use)	

This theory was a positive start to exploring the complexity of influences in assessing credibility. However, research has identified that surface credibility evaluations are influenced by source credibility, content accuracy, and currency (Wierzbicki, 2018). Thus, the steps in the model may not be ordered but rather a fluid and flexible processing of both information sources at the same time, which may rely too heavily on System 1 thinking.

Another model examined the dual roles of motivation and ability in evaluating credibility (Metzger, 2007). The impact of motivation and ability on credibility assessment is empirically supported, with one study showing that people who were motivated to obtain accurate information about a health issue were more likely to initially sift through information using heuristic processing (e.g., assessing surface credibility), before more critically appraising online information (i.e. begin to engage System 2 thinking) (Sillence, Briggs, Harris, & Fishwick, 2007). In another study, Flanagin & Metzger (2000) found that internet users with more experience were more likely to verify online information than less experienced users.

This model recognises that these processes are highly influenced by individual differences and user perceptions (e.g., demographics, experiences, user skills). If an individual is motivated to check credibility but has limited ability to do so (e.g., through unfamiliarity, poor access to the internet), this model proposes that they are more likely to rely on a heuristic evaluation (i.e. System 1 thinking). However, this model extends the MAIN model's focus on heuristic processing by examining an individual's motivation and their ability to evaluate, which may lead to a more systematic and thorough evaluation.

Unifying framework. Derived from empirical research, this model identifies three levels of credibility judgments: construct, heuristic, and interaction (Hilligoss & Rieh, 2008). These three levels are not considered to operate independently, but rather as different judgments at different levels impacting each other.

The construct level describes the user's personal conceptualization of credibility, which provides a particular point of view for judging credibility. The construct level includes concepts such as truthfulness, believability, trustworthiness, objectivity, and reliability. In the associated empirical study, the authors note that trustworthiness was the definition of credibility that participants mentioned most frequently (Hilligoss & Rieh, 2008). The authors also noted that participants used different terms to describe credibility depending on the situation or type of information encountered, indicating that people will adapt their expectations of credibility based on context.

The heuristics level refers to cognitive shortcuts used to estimate credibility, which is akin to using System 1 thinking. In the associated study, participants referred to "convenient" and "quick" assessments, leading to an almost instant judgment of credibility. The study by Hilligoss and Rieh (2008) identified four types of heuristics that people rely on for credibility assessments. The first is media-related heuristics. Participants perceived that books and scholarly journal articles were consistently perceived as credible media, as compared to the internet. As such, media-specific heuristics can serve to increase or decrease credibility concerns and influence the processing of information. The second is source-related heuristics. This was identified as whether the source was familiar or unfamiliar (with familiar sources being perceived as more credible) and whether the source was primary versus secondary (with primary sources being more credible than secondary sources). The third heuristic identified was endorsement-based, whereby participants perceived information to be credible because it was endorsed, recommended, or believed by knowledgeable and trusted individuals. The last heuristic referred to aesthetics, whereby participants used the aesthetic appeal (e.g., how the website looks, how easy it was to navigate) of the online source as a source of credibility.

The interaction level relates to sources or content cues that occur during a specific interaction. The study identified that the interaction level involves three types of interactions: interactions with content cues, source peripheral cues, and information object peripheral cues. Content cues refers to the interaction with the content of the message. Hilligoss and Rieh (2008) found that the primary method by which people interact with content from a credibility assessment perspective was based on personal knowledge, followed by exploring additional sources of the information. Source peripheral cues are those surrounding the online information, such as affiliation, reputation, and type of institution, while peripheral information object cues pertain to the appearance or presentation of the information.

This theory is useful because it considers the user's subjective credibility assessment, as well as heuristic processing, content, and source cues. Furthermore, it considers all this information in the individual context. This theory brings together several factors seen in previous models. However, it does not identify individual or personal differences that impact these assessments, such as technical expertise or the role of relationships.

Aggregated trustworthiness model. This model adds to the literature regarding the role of relationships and social dynamics (Jessen & Jørgensen, 2012). The aggregated trustworthiness model notes that, in the absence of an identified author, people still make credibility judgements about online information in the context of collective judgment, such as 'likes', ratings, or comments (Hargittai, Fullerton, Menchen–Trevino, & Thomas, 2010). The authors propose that the others' feedback plays a role in credibility assessment through three processes: social validation, profiles, and authority/trustee. Social validation means the more people that acknowledge the information, the more likely it is to be perceived as credible. Profiles are the baseline for identity (e.g., social media profiles), while authority and trustee are the known brand or authority. This model shifts from traditional views of expertise and trustworthiness to incorporate online social dynamics. This recognition of the social processes online that are relevant to credibility assessment are evident where social media tools have replaced more traditional authoritative sources.

This model has not been empirically validated, although its concepts have some empirical support (Hargittai et al., 2010; Pettingill, 2006). It remains worthy of consideration due to its introduction of the construct of collective judgment and social processes involved in making a judgment about credibility. These concepts are paramount to credibility assessment in the modern day.

3-S model. The 3-S model encompasses Metzger's (Metzger, 2007) ideas on the importance of motivation and ability in credibility assessment with a focus on trust. In this model, trust comprises of four levels to include individual, interpersonal, relational, and societal (Lucassen & Schraagen, 2011). Trust is defined as, "The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer, Davis, & Schoorman, 1995, p. 712). Previous models of trust have identified factors including disposition to information, relevance, confidence, and willingness to trust (Kelton, Fleischmann & Wallace, 2008). This adds an interesting aspect of online credibility assessment as it pertains to an individual's willingness to take risks regarding credibility and trust.

The 3-S model encompasses two main aspects, information characteristics and user characteristics (Figure 1). Both characteristics impact trust assessments. These two characteristics are akin to Metzger's ability domain (Metzger, 2007). According to this model, when making a credibility judgment, a user will be influenced by the content of the information, how it is presented and the source of the information. After receiving this information, the 3-S model then proposes that there are three different strategies users may apply when judging credibility.

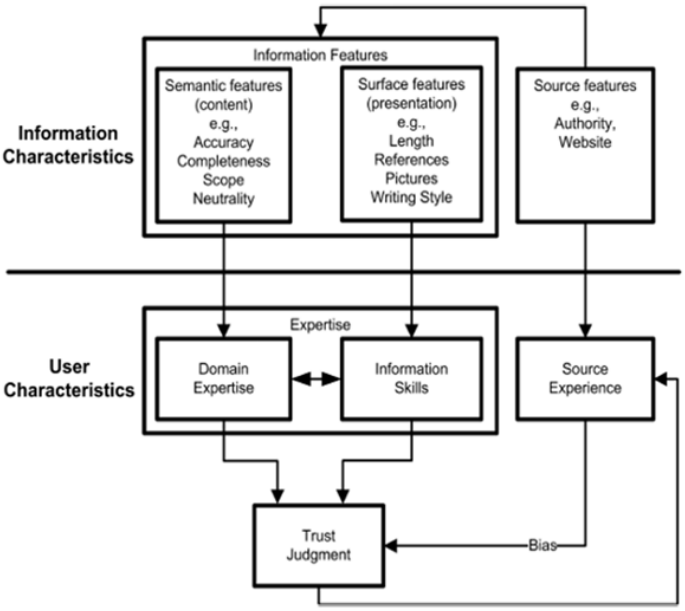

Figure 1. The 3S Model of Trust
Source. Lucaasen & Schraagen (2011)

The user may apply their own expertise through domain expertise or information skills. Research has that experts approach information in their field of expertise differently than novices (Brand-Gruwel, Wopereis, & Vermetten, 2005; Chi, Feltovich, & Glaser, 1981). For example, domain experts are likely to base their judgment primarily on factual accuracy (Lucassen & Schraagen, 2011). In addition to domain expertise

or information skills, the final user characteristic that influences the assessment of trust in this model is source experience. This is commensurate with source-related heuristics in the Unifying Framework (Hilligoss & Rieh, 2008). This model proposes that independent of domain expertise and information skills, source experience can serve to diminish or override System 2 thinking, where users passively rely on their previous experiences with the source.

There is an increase of online material specifically designed to distort, manipulate, or even delude the perceptions of digital consumers with limited empirical examination of cyber deception (Stech, Heckman, Hilliard, & Ballo, 2011). When considering those types of adversarial attempts to thwart online judgment, this model of trust is useful because it highlights several areas where we can explore credibility and trust facets, including both information and user characteristics.

As discussed earlier, while trust does not equate to credibility, credibility is required for trust to be established. This model provides a comprehensive exploration of online factors that may influence a trust judgment and thus credibility. However, this model lacks an explanation of how social dynamics can impact trust judgments online.

Proposed Model for Online Adversary Credibility Assessment

There is no existing model defining online adversary credibility assessment. However, there are common themes and useful constructs in the previously discussed models that can assist us in looking at adversary credibility. This proposed model provides an exploratory method for assessing credibility of cyber actors who have deceptive or malicious intent, which fills a current gap in the literature. Across the models, we have identified three main characteristics - information, user, and interaction characteristics - as shown in Figure 2.

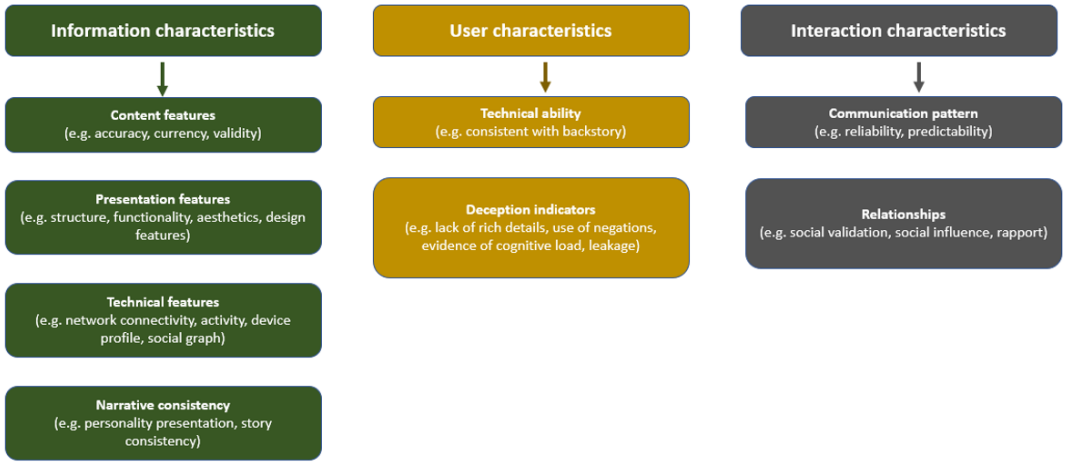

Figure 2. Proposed Model of Online Adversary Credibility Assessment

This paper has explored the scientific evidence for the validity of aspects of credibility and trust models, in order to develop a contemporary, plausible model of assessing adversary credibility. Existing theories of online credibility provide empirical evidence in support of the proposed model, the Online Adversary Credibility Assessment model, as shown in Table 3.

Table 3. Proposed domains of online credibility

Credibility domains	Sub-domains	Model reference
Information characteristics	Semantic features (content)	Lucassen & Schraagen (2011)
	Presentation (surface) features	Fogg (2003)
	Technical features	Metzger (2007)
	Narrative consistency	Wathen & Burkell (2002)
		Hilligoss & Rieh (2008)
		Sundar (2008)
		Jessen & Jorgenson (2012)
Target characteristics	Technical expertise	Fogg (2003)
	Deception indicators	Lucassen & Schraagen (2011)
		Jessen & Jorgenson (2012)
Interaction characteristics	Relationships	Jessen & Jorgenson (2012)
	Communication pattern	Sundar (2008)

Information Characteristics

This model proposes that the first area for credibility assessment for potential adversaries should be in the characteristics of the information being presented. The majority of research and models have identified that information features are important for both System 1 and System 2 processing of credibility information. In this model, we have separated information characteristics into content features (i.e. the actual message), presentation features (i.e. the way the message is presented), technical features (i.e. the technical aspects of the content), and narrative consistency (i.e. how consistent the online information is over time).

Based on a large sample size, Fogg proposed that the source, message, and design are all important in credibility assessment (Fogg, 2003). Overall, they identified that content (e.g., accuracy, comprehensiveness, currency, reliability, and validity) and presentation features (e.g., the structure of the information/ website, technical functionality, aesthetic feel, and interaction design in terms of stability, consistency, and easiness) are fundamental aspects for credibility assessment.

Other information characteristics identified in the literature as demonstrating credibility include the expression of emotion, citing of external sources, strong design features, interactive elements, and level of certainty (Castillo, Mendoza, & Poblete, 2011; Sundar, 2008). Research has also used algorithms to identify credibility features, which include content-based features (e.g., high numbers of unique characters, swear words, pronouns, emojis) and user-based features (e.g., number of followers, length of username) (Alrubaian

et al., 2019; Gupta & Kumaraguru, 2012; Karvonen, 2000). In recent work, the authors identified a number of source and message cues based on a review of empirical studies, outlined in Table 4 (Metzger & Flanagin, 2015).

Table 4. Metzger & Flanagin's (2015) credibility features

Source features	Message features
Professional, attractive page design	Presence of date stamp showing information is current
Easy navigation, well organised site	Citations, links to external authorities
Absence of errors or broken links	Message relevance, tailoring
Certifications, or recommendations, or seals from trusted third parties	Professional-quality and clear writing
Interactive features	Message accuracy, lack of bias, plausibility
Paid access to information	Information breadth and depth
Fast download speed	Description of editorial review process or board
Domain name suffix	
Absence of advertising	
Sponsorship by or links to reputable organisations	
Presence of privacy and security policies	

Technical features can also inform credibility assessment, particularly in the fraud detection literature. Network connectivity refers to attributes that reveal how a profile has connected to the internet and/or uses online platforms. Use of anonymization services (e.g., The Onion Router, proxies, Virtual Private Networks) is sometimes an indicator of suspicious behaviour and may cause additional scrutiny over an account. The Internet Protocol (IP) address associated with an account is also a key network connectivity feature. An account may be considered fraudulent or suspicious if the IP address associated with it was registered using the same registrar typically used by malicious actors or is in the same IP space as known malicious actors (Fukushima, Hori, & Sakurai, 2011; Moura, Sadre, & Pras, 2014). Geographic features paired with an IP also provide clues (e.g., the location of the IP does not match the stated geographic location of the user, the user registers and logs into the service from different IP address locations) (Robert, 2016). Accounts that claim to represent unique individuals but make use of data centre IP space instead of residential IP space are also worthy of additional scrutiny.

Device profile captures attributes associated with the device(s) used to register or access a specific platform; it is analogous to the aggregated trustworthiness model's user profile concept (Jessen & Jørgensen, 2012). Browser fingerprinting is used to extract device profile features (e.g., user agent strings, language packages, screen resolution) from a user's web browser.

Social graph refers to the account's connectivity to other accounts and the evolution of this connectivity over time. Many social graph-based fraud detection approaches rely on the fact that fake accounts generally have fewer connections to real accounts and that fake accounts tend to cluster together. Related to this, it is also often assumed that accounts within a social network tend to be similar, so an account that deviates in appearance from other accounts in a social network it is linked to may be fraudulent (Mislove, Viswanath, Gummadi, & Druschel, 2010).

Observing how a social network forms over time speaks volumes about the authenticity of an account. For example, Facebook considers excessive friend requests suspicious (Robert, 2016). Research also suggests fake accounts (and in particular, automated accounts e.g., "bots") spend their time much differently than authentic accounts; for example, fake accounts spend more time building new connections than interacting with existing connections (Wang, Konolige, Wilson, Wang, Zheng, & Zhao, 2013; Yang, Wilson, Wang, Gao, Zhao, & Dai, 2011).

Activity-based indicators include attributes related to the nature and frequency of activity on a single account or coordinated activities across multiple accounts. This includes temporal patterns; times of day an account is active, the frequency with which an account engages with content or other accounts on a platform, and the frequency with which an account creates/shares/ posts content. One study, using the Chinese social network Renren, found the number of sessions per user, number of sessions through the day, average session length per user, average number of clicks per user, and average time interval between clicks per session per user were all indicators of fake accounts (Yang et al., 2011).

Click analysis, typing speeds, and mouse movements are activity-based fraud indicators known as 'behavioral biometrics', such as Google's 'invisible' CAPTCHA (Awad, 2017; Bo, Zhang, Li, Huang, & Wang, 2013). In one study, multiple accounts/identities managed by the same person were identified using features such as site familiarity and time taken to make posts (Tsikerdekis & Zeadally, 2014).

Campaigns of suspicious or manipulative activity can be detected by looking for coordinated activity or similar behavior across multiple accounts. For example, in one study authors uncovered over two million malicious accounts and 1,156 campaigns on Facebook and Instagram by identifying accounts that acted similarly at similar times in a sustained window of time (Cao, Yang, Yu, & Palow, 2014). As another example, researchers looking for coordinated patterns of activity on Twitter and various other platforms have identified many Russian-sponsored troll accounts (Jamieson, 2018).

A growing area of research in the fraud detection space related to user profiles is fake image detection. Companies such as Facebook and Airbnb have already employed various fake image detection techniques such as image hashing to detect the re-use of existing images. However, the rise of convincing "deep fake" image technology has prompted increased interest in developing statistical techniques to detect fake images used in online profiles (Menotti et al., 2015).

Narrative consistency is particularly important in adversary credibility assessment. In the proposed model, this is defined as the consistency of an online persona and online information over time. It includes both story consistency and personality presentation. Unfortunately, in more general models of online information credibility, narrative consistency has been neglected as a topic of research. The importance of narrative consistency, particularly in examining individual online accounts, is particularly relevant for heuristic processing. Research has shown that people expect consistency in presentations, and consistent presentations have greater influence over others than online presentations that are inconsistent (Isbister & Nass, 2000). When online contacts behave in an expected and consistent way, we are much less likely to apply System 2 thinking in a credibility assessment. In this model, we propose that examinations of narrative consistency in online information and profiles are essential in adversary credibility assessment.

The importance of content, technical features, and presentation are evidenced in the majority of the models discussed, such as modality in the MAIN model, source and surface characteristics in the dual processing models and information characteristics in the 3-S model. The growing fraud detection literature can help us identify additional technical indicators to support credibility assessment. Further examination of the validity of narrative consistency as a tool for detecting fake or adversary online personas is warranted.

User Characteristics

Trustworthiness, expertise, dynamism, and attractiveness of the communicator are all relevant to judge overall credibility of the message (Fogg, 2003). In one study, researchers explored what characteristics shaped people's perception of credibility on Twitter. The user's influence, reputation, and topical expertise (judged from the user's biography) resulted in enhanced perceived credibility (Morris, Counts, Roseway, Hoff, & Schwarz, 2012). In this proposed model of online adversary credibility, user characteristics are described as aspects of the online user (e.g., a potential adversary) that can indicate credibility concerns. This includes deception indicators and technical ability. Much of the previous discussion of fraud detection literature is pertinent here, but also the consistency of the technical ability with stated skills, backstory, and account settings are important to consider.

Deception indicators suggest intentional and misleading efforts to create a false belief in another. While such indicators are not specifically identified in the theoretical literature, trustworthiness and credibility are reliant on an honest or reliable relationship. Indicators of deception are contra-indicative to credibility. Digital deception requires manipulation enacted through a technologically enabled message (Donath, 1999; Hancock, 2012). However, deception detection is notoriously difficult and, on average, leads to an accuracy rate of not much greater than chance of around 55% (Vrij, 2000). This is because there are no obvious clues to deceit and the differences between liars and truth tellers are subtle.

Deception can be detected using both verbal and nonverbal cues. The empirically supported cognitive assessment approach to detecting deception uses cognitive techniques to elicit and enhance potential differences between truth tellers and liars. In a meta-analysis, this approach had accuracy rates of 71% versus 56% in a standard approach (Vrij, Fisher, & Blank, 2017). Through the cognitive assessment approach, research has identified a number of potential verbal cues including evidence of cognitive load, leakage, evidence of poor coping, inconsistent behavior, an absence of verifiable details, and lack of detail (Jupe, Leal, Vrij, & Nahari, 2017; Nahari, Vrij, & Fisher, 2012; Vrij et al., 2017). Nonverbal cues identified in the literature to differentiate between truth tellers and liars are related to methods such as working memory, attention, arousal, motivation, and self-presentation (Arciuli, Mallard, & Villar, 2010).

A language behavior approach assumes that a liar's language reflects their emotions and cognitions (Toma & Hancock, 2010). Research on automated textual analysis suggests that there are detectable differences in linguistic patterns in communication that is true versus deceptive (Zhou & Zhang, 2008). In fact, some studies have shown that linguistic cues in deception can increase the ability to detect deception to 67% (Newman, Pennebaker, Berry, & Richards, 2003). Potential linguistic indicators of deception include a lack of rich details, reduced self-reference (e.g., 'I' statements), increased use of negations (e.g., 'no', 'not', 'never'), being overly specific, having a lower number of named entities, and using fewer evaluations and judgments (Hauch, Blandón-Gitlin, Masip, & Sporer, 2015; Kleinberg, van der Toolen, Vrij, Arntz, & Verschuere, 2018; Masip, Sporer, Garrido & Herrero, 2005; Toma & Hancock, 2010).

Credibility may also be assessed by whether or not the stated technical expertise is commensurate with the back story and demonstrated technical ability. Following the construct of narrative consistency, this technical consistency is the demonstration that the online persona is as technically savvy as he or she proports. Claims of expertise in coding acumen, debugging techniques, or other technical skills can be compared to the cyber security practices they maintain as well as examples of their technical writing or any technical processes or procedures they discuss, publish, opine on, or demonstrate. The assessment of technical narrative characteristics naturally engages System 2 thinking, which is essential in determining credibility of potential adversaries or misleading personas. Thus deception indicators and inconsistencies in claims of technical expertise as compared to their lack of demonstrated technical expertise can highlight credibility concerns.

Interaction Characteristics

Trust plays an important role in determining credibility within relationships. Credibility is intertwined with the concept of trust and is an inherent characteristic of meaningful and valuable social interaction (Cugelman, Thelwall, & Dawes, 2008; du Plessis, Angelopulo, & du Plessis, 2006; Fogg & Tseng, 1999). McKnight and Chervany created an interdisciplinary model of trust, which proposes that the many definitions of trust fall into different conceptual types: behaviors, beliefs, attitudes, and dispositions (McKnight & Chervany, 2000). They identified 16 trust-related characteristics, which were distilled to five conceptual categories of characteristics: benevolence, competence, predictability, integrity, and other (openness, carefulness, attractiveness, and shared understanding).

The aggregated trustworthiness model recognises how online credibility is influenced by social processes, such as followers and likes. The notion of social proofing is not new in influence psychology. Social validation is a similar concept – the more people that acknowledge information, the more likely it is to be perceived as credible (Cialdini, 2009; Metzger & Flanagin, 2015). Research has focussed on social online processes that influence credibility assessment. While traditionally authority figures provided opinion to influence credibility, experiential credibility means that people can now provide credible information on the basis of experience rather than authority (Pure et al., 2012). For example, customer reviews have been found to be stronger indicators of trustworthiness than store reputation (Utz, Kerkhof, & Van Den Bos, 2012). Informational social influence is the tendency to "accept information obtained from another as evidence about reality" (Deutsch & Gerard, 1955, p. 629). Therefore, informational social influence is a means of gaining information under circumstances where people are uncertain about their own perceptions. This is similar to Word of Mouth (Westbrook, 1987). Complimenting this is referent social influence, which works to inform opinion and assessment through group identity – people look to others to guide their own opinions within their group norms (Turner, Wetherell, & Hogg, 1989). These concepts may play an influential role in credibility.

There are trust-enhancing factors in the online environment, which users interpret when making trust-related decisions. These factors fall into several categories, including technical factors and human behavior-related factors. Technological factors may include the design, language, and reliability of a website or online forum. For example, people are attracted to a smooth, flawless website design. If they notice flaws, such as typos, it can adversely affect their decision to trust that website (Jones & Moncur, 2018). These factors are similar to the surface features of the 3S model and information characteristics mentioned above.

Additionally, human behavior factors are significant to the trust and relationship formation process. These factors may include establishing rapport, transaction reliability, and customer service. They are similar to interaction characteristics previously discussed, such as benevolence, competence, predictability, and integrity. Being able to identify credible relationships and valid communication patterns may serve to enhance our confidence in assessment. Conversely, there are many relationship and communication factors that may indicate potential malicious activity.

Discussion

The evolving online credibility literature and research can be applied to adversary credibility assessment online. Identifying cyber adversaries is not just the business of intelligence agencies anymore. In fact, the responsibility lies with governments, social media companies, and, most importantly, the individual. Factors identified as influencing decision-making and judgment regarding credibility, deception, and trust are useful to determine areas for assessing adversary credibility. Note that attributes of credibility should not be considered in isolation or without recognition of the context. An attribute becomes an indicator of credibility concerns when it is incongruous with other attributes or deviates from the observed norm. In turn, being able to identify what "normal" looks like is essential in credibility assessment.

The Online Adversary Credibility Assessment model proposed in this paper for assessing the credibility of potential adversarial and fake personas provides a broader, more comprehensive perspective of the individual being assessed incorporating technical, personal, communicative, and interactive characteristics identified in the current empirical body of literature. The proposed model also requires the engagement of System 2 thinking to assess what looks real and to determine what could be fake or misleading. The consequences of fake or misleading persona-produced material is paramount, as has been illustrated with the impact of false information regarding the 2016 U.S. presidential election campaigns. The Online Adversary Credibility Assessment model highlights potential opportunities for public engagement and education about the risk associated with the online engagement of cyber adversaries. There are potential human and technical advances that could be considered on the basis of this model. This model could be adapted for education purposes for the general public, as well as utilized by organizations and governments in monitoring and assessing cyber actors and foreign influence. Further, this model may have value in the development of cyber adversary detection so that social media and similar online applications can incorporate this model into their authentication processes to assist in identifying potential false accounts, personas, or material.

Limitations and Future Directions

In this paper, we have used existing literature on online credibility to examine aspects of credibility assessment as applied to adversary assessment. While we have tried to synthesise the literature, this is not a systematic review and there may be literature or studies that we have not included in this review. A systematic review would enable a comprehensive understanding of the full research regarding online credibility and how it relates to adversary assessment.

This research has informed the proposal of a model for assessing online adversary credibility but ongoing evaluation as to the reliability and validity of these domains and factors is essential. Though the proposed adversary credibility model remains theoretical, it is rooted in empirical literature and provides the basis for future studies to explore and examine the empirical support for this model. Testing the empirical validity of

this model could be achieved through an examination of the model's ability to identify and predict malicious cyber actors based on online information.

Summary

Being able to assess online credibility has become everybody's business. People are continuously making judgments about the credibility of online information, but research has shown that we often rely on heuristics to make decisions in situations that should require more thoughtful consideration (Kahneman & Frederick, 2005). How we direct those considerations needs to be guided by science. This paper provides an overview of research that is then used by the authors to develop a theoretical credibility model of malicious cyber actors. This model has three empirically driven areas for assessment; information, user, and interaction characteristics. Information characteristics explores the online content, presentation, consistency, and technical features. This model then highlights the importance of the user in credibility assessment and identifies exploring both technical ability and deception indicators. Lastly, this model identifies interaction characteristics as relevant to credibility assessment through exploring communication patterns and relationships. It is intended to provide a comprehensive spectrum of considerations when assessing the veracity of online actors and sources. Future research should aim to examine the empirical validity of this proposed model.

Declaration of Interest Statement

The authors declare that they have no conflicts of interest.

References

Alrubaian, M., Al-Qurishi, M., Alamri, A., Al-Rakhami, M., Hassan, M. M., & Fortino, G. (2019). Credibility in online social networks: A survey. *IEEE Access, 7*. doi: 10.1109/ACCESS.2018.2886314

Arciuli, J., Mallard, D., & Villar, G. (2010). "Um, I can tell you're lying": Linguistic markers of deception \ versus truth-telling in speech. *Applied Psycholinguistics, 31*, 397–411. doi: 10.1017/S0142716410000044

Australian Cyber Security Centre. (2020). ACSC Annual Cyber Threat Report July 2019 to June 2020. Retrieved July 7, 2020, from https://www.cyber.gov.au/sites/default/files/2020-09/ACSC-Annual-Cyber-Threat-Report-2019-20.pdf

Awad, A. (2017). Collective framework for fraud detection using behavioral biometrics. In I. Traore, A. Awad, & I. Woungang (Eds.), *Information Security Practices: Emerging Threats and Perspectives* (pp. 29–37). Springer. doi: 10.1007/978-3-319-48947-6_3

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*(4), 446–458. doi: 10.1037/0096-3445.121.4.446

Bo, C., Zhang, L., Li, X. Y., Huang, Q., & Wang, Y. (2013). SilentSense: Silent user identification via touch and movement behavioral biometrics. *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM,* 187–190. doi: 10.1145/2500423.2504572

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

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior, 21*(3), 487–508. doi: 10.1016/j.chb.2004.10.005

Cao, Q., Yang, X., Yu, J., & Palow, C. (2014). Uncovering large groups of active malicious accounts in online social networks. *Proceedings of the ACM Conference on Computer and Communications Security*, 477–488. doi: 10.1145/2660267.2660269

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW 2011. doi: 10.1145/1963405.1963500

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152. doi: 10.1207/s15516709cog0502_2

Cialdini, R. (2009). *Influence: The Psychology of Persuasion*. New York, NY: Harper Collins.

Clement, J. (2020). *Internet usage worldwide – statistics and facts*. Retrieved from https://www.statista.com/topics/1145/internet-usage-worldwide/

Cugelman, B., Thelwall, M., & Dawes, P. (2008). Website credibility, active trust and behavioural intent. *Proceedings of 3rd International Confererence Persuasive Technology*, 47–57. doi: 10.1007/978-3-540-68504-3-5

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology, 51*, 629–636. doi: 10.1037/h0046408

Donath, J. (1999). Identity and deception in the virtual community. In M. A. Smith & P. Kollock (Eds.), *Communities in Cyberspace* (pp. 29–59). Abingdon, UK: Routledge/Taylor & Francis Group. doi: 10.1519/JSC.0b013e3181e4f7a9

du Plessis, C., Angelopulo, G., & du Plessis, D. (2006). A conceptual framework of corporate online communication: A marketing public relations (mpr) perspective. *Communicatio, 32*(2), 241–263. doi: 10.1080/02500160608537972

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association, 43*(2), 97–116. doi: 10.1080/23808985.2019.1602782

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism and Mass Communication Quaterly, 77*(3), 515–540. doi: 10.1177/107769900007700304

Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. *Conference on Human Factors in Computing Systems - Proceedings*, 722–723. doi: 10.1145/765891.765951

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. *Conference on Human Factors in Computing Systems - Proceedings*, 80–87. doi: 10.1145/302979.303001

Fukushima, Y., Hori, Y., & Sakurai, K. (2011). Proactive blacklisting for malicious web sites by reputation evaluation based on domain and IP address registration. *Proc. 10th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications, TrustCom 2011, 8th IEEE Int. Conf. on Embedded Software and Systems, ICESS 2011, 6th Int. Conf. on FCST 2011*. doi: 10.1109/TrustCom.2011.46

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. *ACM International Conference Proceeding Series*. doi: 10.1145/2185354.2185356

Hancock, J. T. (2012). Digital deception: Why, when and how people lie online. In A. N. Joinson, T. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford Handbook of Internet Psychology* (pp. 287–301). Oxford, UK: Oxford University Press. doi: 10.1093/oxfordhb/9780199561803.013.0019

Hargittai, E., Fullerton, L., Menchen–Trevino, E., & Thomas, K. Y. (2010). Trust Online: Young adults' evaluation of web content. *International Journal of Communication, 4*, 468–494. Retrieved July 7, 2020, from http://ijoc.org.

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. doi: 10.1177/1088868314556539

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing and Management, 44*, 1467–1484. doi: 10.1016/j.ipm.2007.10.001

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and Persuasion: Psychological Studies of Opinion Change*. Yale, CT: Yale University Press. doi: 10.2307/2087772

Igartua, J. J., & Cheng, L. (2009). Moderating effect of group cue while processing news on immigration: Is the framing effect a heuristic process? *Journal of Communication, 59*(4), 726–749. doi: 10.1111/j.1460-2466.2009.01454.x

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies, 53*(2), 251–267. doi: 10.1006/ijhc.2000.0368

Jamieson, K. H. (2018). *Cyber-War*. Oxford, UK: Oxford University Press.

Jessen, J., & Jørgensen, A. H. (2012). Aggregated trustworthiness: Redefining online credibility through social validation. *First Monday, 17*(1). doi: 10.5210/fm.v17i1.3731

Jones, H. S., & Moncur, W. (2018). The role of psychology in understanding online trust. In *Psychological and Behavioral Examinations in Cyber Security* (pp. 109–132). Hershey, PA: IGI Global.

Jupe, L. M., Leal, S., Vrij, A., & Nahari, G. (2017). Applying the verifiability approach in an international airport setting. *Psychology, Crime and Law, 23*(8), 812–825. doi: 10.1080/1068316X.2017.1327584

Karvonen, K. (2000). The beauty of simplicity. *Proceedings of the Conference on Universal Usability,* 85–90. doi: 10.1145/2133806.2133808

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology, 59*(3), 363–374. doi: 10.1002/asi.20722

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology, 32*(3), 354–366. doi: 10.1002/acp.3407

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication, 50*, 46–70. doi: 10.1111/j.1460-2466.2000.tb02833.x

Langenderfer, J., & Shimp, T. A. (2001). Consumer vulnerability to scams, swindles, and fraud: A new theory of visceral influences on persuasion. *Psychology and Marketing, 18*(7), 763–783. doi: 10.1002/mar.1029

Lucassen, T., & Schraagen, J. M. (2011). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology, 62*(7), 1232–1242. doi: 10.1002/asi.21545

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime and Law, 11*(1), 99–112. doi: 10.1080/10683160410001726356

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. doi: 10.5465/amr.1995.9508080335

McKnight, D. H., & Chervany, N. L. (2000). What is trust ? A conceptual analysis and an interdisciplinary model. *Proceedings of the 2000 Americas Conference on Information Systems AMCI2000 AIS*, 382. Retrieved July 7, 2020, from https://aisel.aisnet.org/amcis2000/382/

Menotti, D., Chiachia, G., Pinto, A., Schwartz, W. R., Pedrini, H., Falcão, A. X… & Rocha, A. (2015). Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security, 10*(4), 864–879. doi: 10.1109/TIFS.2015.2398817

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58*(13), 2078–2091. doi: 10.1002/asi.20672

Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment online. In S. Sundar (Ed.), *The Handbook of the Psychology of Communication Technology* (pp. 445–466). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118426456.ch20

Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010). You are who you know: Inferring user profiles in online social networks. *WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. doi: 10.1145/1718487.1718519

Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? Understanding microblog credibility perceptions. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 441–450. doi: 10.1145/2145204.2145274

Moura, G. C. M., Sadre, R., & Pras, A. (2014). Bad neighborhoods on the internet. *IEEE Communications Magazine, 52*(7), 132–139. doi: 10.1109/MCOM.2014.6852094

Mueller, R., Helderman, R. S., & Zapotosky, M. (2019). Report on the Investigation into Russian Interference in the 2016 Presidential Election. Retrieved from https://www.justice.gov/storage/report.pdf

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Does the truth come out in the writing? Scan as a lie detection tool. *Law and Human Behavior, 36*(1), 68–76. doi: 10.1037/h0093965

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*, 665–675. doi: 10.1177/0146167203029005010

Pettingill, L. (2006). Trust without knowledge: How young person's carry out research on the Internet. In *GoodWork Project Report Series*. Retrieved July 7, 2020, from http://pzweb.harvard.edu/ebookstore/pdfs/goodwork48.pdf

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. doi: 10.1111/j.1559-1816.2004.tb02547.x

Pure, R. A., Markov, A. R., Mangus, J. M., Metzger, M. J., Flanagin, A. J., & Hartsell, E. H. (2012). Understanding and evaluating source expertise in an evolving media environment. In T. Takseva (Ed.), *Social Software and the Evolution of User Expertise: Future Trends in Knowledge Creation and Dissemination* (pp. 37–51). Hershey, PA: IGI Global. doi: 10.4018/978-1-4666-2178-7.ch003

Purplesec. (2020). *2020 Cyber Security Statistics. The Ultimate List Of Stats, Data & Trends*. Retrieved June 20, 2020, from https://purplesec.us/resources/cyber-security-statistics

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology, 41*(1), 307–364. doi: 10.1002/aris.2007.1440410114

Robert, H. (2016). *Spam Fighting @ Scale 2016*. Retrieved from https://code.fb.com/security/spam-fighting-scale-2016/

Sillence, E., Briggs, P., Harris, P. R., & Fishwick, L. (2007). How do patients evaluate and make use of online health information? *Social Science and Medicine, 644*, 1853–1862. doi: 10.1016/j.socscimed.2007.01.012

Stech, F., Heckman, K. E., Hilliard, P., & Ballo, J. R. (2011). Scientometrics of deception, counter-deception, and deception detection in cyber-space. *PsychNology Journal, 9*(2), 79–122.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital Media, Youth, and Credibility* (pp. 73–100). The MIT Press. doi: 10.1162/dmal.9780262562324.073

Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology, 58*, 366–378. doi: 10.1002/asi.20511

Talboom, S., & Pierson, J. (2013). Understanding trust within online discussion boards: Trust formation in the absence of reputation systems. In *IFIP Advances in Information and Communication Technology* (pp. 83–99). doi: 10.1007/978-3-642-38323-6_6

Taylor, R. S. (1986). Value-added processes in the information life cycle. In *Journal of the American Society for Information Science*. Ablex. doi: 10.1002/asi.4630330517

Toma, C. L., & Hancock, J. T. (2010). Reading between the lines: Linguistic cues to deception in online dating profiles. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. doi: 10.1145/1718918.1718921

Tsikerdekis, M., & Zeadally, S. (2014). Multiple account identity deception detection in social media using nonverbal behavior. *IEEE Transactions on Information Forensics and Security, 9*(8), 1311–1321. doi: 10.1109/TIFS.2014.2332820

Turner, J. C., Wetherell, M. S., & Hogg, M. A. (1989). Referent informational influence and group polarization. *British Journal of Social Psychology, 28*(2), 135–147. doi: 10.1111/j.2044-8309.1989.tb00855.x

Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: Heuristics and biases. Science, 185*, 1124–1131. doi: 10.1126/science.185.4157.1124

Twitter Public Policy. (2018). *Update on Twitter's review of the 2016 US election*. Retrieved June 20, 2020, from https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html

Utz, S., Kerkhof, P., & Van Den Bos, J. (2012). Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. *Electronic Commerce Research and Applications, 11*(1), 49–58. doi: 10.1016/j.elerap.2011.07.010

Vishwanath, A. (2015). Diffusion of deception in social media: Social contagion effects and its antecedents. *Information Systems Frontiers, 17*, 1353–1367. doi: 10.1007/s10796-014-9509-2

Vishwanath, A., Harrison, B., & Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research, 45*(8), 1146–1166. doi: 10.1177/0093650215627483

Vrij, A. (2000). *Detecting Lies and Deceit: The psychology of lying and its implications for professional practice*. Hoboken, NJ: John Wiley and Sons.

Vishwanath, A., Harrison, B., & Ng, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology, 22*(1), 1–21. doi: 10.1111/lcrp.12088

Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., & Zhao, B. Y. (2013). You are how you click: Clickstream analysis for sybil detection. *Proceedings of the 22nd USENIX Security Symposium*, 241–256. Retrieved June 20, 2020, from https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_wang_0.pdf

Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 134–144. doi: 10.1002/asi.10016

Westbrook, R. A. (1987). Product/consumption-based affective responses and post-purchase process. *Journal of Marketing Research, 24*, 258–270. doi: 10.2307/3151636

Wierzbicki, A. (2018). *Web Content Credibility. Springer International Publishing*. doi: 10.1007/978-3-319-77794-8_1

Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., & Dai, Y. (2011). Uncovering social network sybils in the wild. *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2. doi: 10.1145/2068816.2068841

Zhou, L., & Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM, 51*(9), 119–122. doi: 10.1145/1378727.1389972