

Lifting the lid on manipulative website contents: A framework mapping contextual and informational feature combinations against associated social cognitive vulnerabilities

Piers MacLean  | Marie Cahillane  | Victoria Smy

Applied Psychology Group, Centre for Electronic Warfare, Information and Cyber, Cranfield University, Defence Academy of the United Kingdom, Shrivenham, UK

Correspondence

Marie Cahillane.

Email: m.cahillane@cranfield.ac.uk

Funding information

Defence Science and Technology Laboratory, Grant/Award Number: DSTLX1000133521

Abstract

Website contents can be designed to influence individual and group decision-making for social, political or financial gain. A novel working theoretical framework was developed to provide insights into where website contents have been designed to exploit common cognitive vulnerabilities (CVs) amongst audiences; a form of social cognitive hacking. A literature synthesis on CVs, website content design and credibility identified features on a context credibility (CC) dimension that were mapped against features on an information validity (IV) dimension. Alignment of CC and IV feature pairings with CVs enables evaluation and interpretation of nuanced website content for identification of attempted influence undetectable by machine algorithms. Subjective responses to prompts about features on each dimension generate pairs of numerical values. The value pairs indicate within a quad-graph the possible presence and extent of CV exploitation. Each value pair can be traced to the underlying CC-IV features and CVs being targeted. External prima facie assessment suggests the framework can form the basis for standardised human analysis of website contents. Application of the framework provides insights into what and how feature combinations

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. Social and Personality Psychology Compass published by John Wiley & Sons Ltd.

may be manipulated for CV exploitation. It has potential for application in fields such as intelligence analysis, education, and marketing.

KEYWORDS

cognitive vulnerabilities, context credibility, information validity, online cognitive hacking

JEL CLASSIFICATION

Social Cognition

1 | INTRODUCTION

The amount of information available online continues to grow. It is fed by website content from mainstream online publishers, organisations and individuals as well social media platform content. Anyone with a minimum of technical skill and knowledge can publish content online using 'out-of-the-box' website building services. The average consumer of online information may lack the knowledge and experience to engage with it critically to assess its quality and veracity (Doty, 2015; Pennycook & Rand, 2019). Even if engaging with content analytically, their reasons for doing so in the first place introduce bias into how the information is interpreted as a result of motivated reasoning (Dhoju et al., 2019; Soon & Goh, 2018; Tsang, 2022). These are recurrent phenomena alongside a frequent lack of website editorial oversight and quality control measures leading to the proliferation of incomplete, out of date, inaccurate or out of context material. Moreover, source information regarding the provenance of content may be unavailable, masked or difficult to interpret. These issues are compounded by the distortion of facts and meaning resulting from the aggregation of information and media content from multiple sources and irrelevant contexts (Metzger & Flanagin, 2013; Rieh & Danielson, 2007). Such factors create fertile ground for the online dissemination of disinformation and propagation of misinformation. Indeed, the discernible spread of online dis- and mis-information as a means to skew individuals' worldviews, destabilise democracies and disrupt economic development is more apparent following recent conflicts (Helmus et al., 2018). It also became a threat to social and political stability during the COVID-19 public health crisis (Vidgen et al., 2021). To understand how such threats arise it is necessary to examine the factors contributing towards individual and, by extension, social cognitive vulnerability to dis- and mis-information (Pantazi et al., 2021).

As will be shown, cognitive vulnerabilities (CVs) stem from the mind's use of heuristics that is, mental shortcuts, and biases in its subconscious information processing. For example, perceptions of expertise or positive reputation are likely to influence whether or not people engage with or come to rely upon a website, source or author (e.g. Metzger et al., 2010; Metzger & Flanagin, 2013). This subconscious bias towards perceived expertise or reputation represents a CV because perceptions can be intentionally manipulated by producers of websites and their contents (Cybenko et al., 2002). (See Supporting Information S1: Appendix A for further CV examples.) These vulnerabilities in cognition may be exploited for 'cognitive hacking' of thought processes that is, the intentional manipulation of how consumers of information perceive informational content (Cybenko et al., 2002). Whole societies as well as individuals are vulnerable (Pantazi et al., 2021) therefore, when social groupings are targeted by such techniques we refer to this as 'social cognitive hacking'. Although machine algorithms may be capable of detecting overt threats from social cognitive hacking within content (Zhang et al., 2018; Zhou & Zhang, 2008), some attempts are so subtle that human scrutiny is required (Marcellino et al., 2017).

Media message credibility may be influenced by non-source factors, such as the medium or channel of delivery and even the structure of the messages themselves (Metzger et al., 2003), leading to calls for a separation of the

concept by situating it at three different loci: source, medium, and message (Appelman & Sundar, 2016). In terms of 'medium' or what we will refer to as 'context credibility (CC)', users of information base credibility judgements about websites on features such as design look, site structure, and usefulness of information (Harris, 2020; Metzger et al., 2010; Metzger & Flanagin, 2013; Warnick, 2004). These judgements are frequently underpinned by the role of heuristics (Metzger, 2007; Metzger et al., 2010; Metzger & Flanagin, 2013). Many educational institutions draw upon guidance to teach students how to evaluate and assess the validity of different information types (Harris, 2020).

This paper addresses the current gap in understanding how exploitable CVs relate to the contextual credibility and informational validity of online content. It achieves this by drawing upon the science of credibility evaluation, frameworks for information validity (IV) assessment, and website user interface design guidance to highlight their implications for CVs. We take the view that heuristics and biases constitute inherent vulnerabilities in human cognition that are generic to human society. Therefore, vulnerabilities need to be understood in terms of social cognition and how some social actors attempt to manipulate social groupings through designing online content to exploit collective vulnerabilities. A novel working theoretical framework which brings contextual credibility, informational validity and CVs together is presented. It provides a basis for further research that will have implications for different domains for example, intelligence analysis, education, and marketing. It will help develop insights into how website contents can be designed intentionally to influence individual and group decision-making and subsequent behaviours for social, political, or financial gain.

2 | BACKGROUND LITERATURE

This section discusses the background literature relating to the inductive development of the working theoretical framework in Section 3.

2.1 | Disinformation and misinformation

Disinformation represents the deliberate creation and spreading of false and/or manipulated information to mislead and deceive others (House of Commons et al., 2019; Martens et al., 2018; Tandoc et al., 2018). It often aims to connect with any prevailing public discontent or widely held negative attitudes. It does so using fabricated information to amplify reductive social interpretations that confirm existing beliefs, support desired conclusions, or prompt certain strong emotions regarding social groups and events (Grace & Leskovich, 2015). Fabricated or manipulated information is often ideologically fluid, where content can overlap with different oppositional political groups, promoting social division (Paul & Matthews, 2016; Samoilenko, 2018). Ultimately, disinformation focuses on degrading the decision-making capabilities of information consumers through non-rational means (e.g. deception), targeting perception and cognition (Arif et al., 2018). In contrast, misinformation generally connotes the inadvertent sharing of false, inaccurate, or contextually misplaced information (Georgiadou et al., 2018). Dis- and misinformation are inputs to human information processing and the decision-making processes underpinning many aspects of human attitude formation and behaviour.

2.2 | Information credibility judgements

An output of human information processing of online content is a judgement regarding its credibility. Judgements are often based on the information consumer's initial perception and early cognitive processing rather than later, more systematic, consideration of the information required to evaluate quality and veracity (Lee, 2018). Dual

processing models of credibility evaluation offer insights as to how information consumers process what they see and hear online (Ecker et al., 2014; Metzger, 2007). A major principle of the dual processing theory is that people sense-make via two mental processing routes (Kahneman, 2011): (i) unconscious and heuristic that is, reliant upon the application of mental shortcuts; and (ii) conscious and systematic, requiring deliberate analytical thinking. The extent to which heuristics are triggered is determined by the depth of cognitive processing applied to what is seen and heard. Kahneman (2011) describes these two routes as System 1 and System 2 thinking.

System 1 occurs automatically, with minimal effort and no voluntary control, whilst System 2 requires effortful, complex, and systematic mental activities. Both interact and System 2 can change the operations of System 1 by 'programming the autonomic functions of attention and memory' (Kahneman, 2011, p. 23). System 1 generates suggestions which become impressions for System 2 thereby framing conscious analytical thought. If endorsed by System 2, they become the sources of beliefs and actions. The decisional outcomes of information processing and subsequent behaviours will largely be influenced by the interactions between System 1 and System 2 and whichever has been predominant.

Heuristic information processing is a common way humans cope with uncertainty and information overload (Martens et al., 2018; Metzger et al., 2010; Metzger & Flanagin, 2013; Samoilenko, 2018; Sundar, 2008; Zhang et al., 2018). People often lack the motivation to monitor their decision-making processes. Even when people are motivated, achieving accurate awareness of their decision processes is difficult (Lerner et al., 2015). Individuals often apply 'filters' to information that may have little to do with logic and reasoning (Niehoff & Shah, 2017). Dual processing models suggest that when motivation or ability to assess information quality and credibility is low, it is more likely that individuals will unconsciously rely on simple 'peripheral credibility features' in the informational context instead of consciously and systematically the information itself. In contrast, when motivation and ability is high, individuals are more likely to adopt conscious-systematic processing to determine information credibility (Metzger et al., 2010). Thus, the processing route adopted and the degree of cognitive effort required to form judgements are determined by the relevance of the information to the intended audience and their level of involvement (Bellur & Sundar, 2014).

An alternative to the dual processing perspective is the limited capacity model that is, the finite mental resources available to individuals for processing all features within information. In the face of online information overload, only the most salient features within the content may be attended to (Lang, 2000) and heuristic processes are more likely to be triggered. Further explanations for information processing limitations are proposed by theories of bounded rationality and satisficing. People do not always demonstrate perfect rationality due to limited cognitive resource (Soon & Goh, 2018) and subconsciously minimise cognitive effort to strike a balance between effort and efficient outcomes that is, satisfice (Pirolli, 2005). Consequently, people will focus upon easily accessible and understandable information to economise on the level of cognitive effort required. Belief is cognitively easier than disbelief; to identify and handle dis- and mis-information an individual must first acquire a proposition before they can discredit or accept it, which requires more cognitive effort (Friedman, 2017).

Heuristic thinking may be determined by attitudes, beliefs and social norms which introduces bias into information processing. Consequently, heuristic-based processing of information can be unreliable and lead to erroneous outcomes and as such, it represents a vulnerability within human cognitive processing. Although insights into such CVs may be exploited for positive outcomes such as altering negative attitudes towards climate action (Soon & Goh, 2018), they may also be exploited negatively by malicious actors (state or non-state) through the manipulation of nuanced content. Such exploitation characterises 'cognitive hacking' of individuals' subconscious thought processes to disrupt, divide, or confuse (Cybenko et al., 2002). The ultimate goal is behaviour modification, even if how this is achieved is not fully understood by malicious content creators (Bastick, 2021). Cognitive hacking also affects the decision-making of the social groups to which individuals belong and their subsequent behaviours and actions. It may be carried out for social (disruption), political (ideological) or financial gain (Tandoc et al., 2018; Waltzman, 2017).

2.3 | Influence of design on information interpretation

In current online information contexts, the predominant processing activities required for sense-making of perceptual inputs are found in the visual and auditory processing centres of the brain, which receive and interpret what might be thought of as 'data' as consumers of content focus their attention on the text, sounds, and visual material they encounter. Human perception and interpretation of information is shaped by how known and proven facts, as well as unproven ones, are organised and presented. Facts, organised into forms and patterns, become information and can be moved around into different patterns 'to conceal or reveal what the informer intends' (Shedroff, 2001, p. 28). The way in which information is organised creates or shapes the meaning inferred by the receiver of the information.

In the context of the current research, the features of a website are considered to be those characteristics that will, without a conscious attempt to understand their meaning, affect perception and understanding of the website and its content. For example, as with factual political news, political disinformation can also trigger a strong affective response that is, anger (Barfar, 2019). The difference is that affective responses resulting from political disinformation are intended to influence how the information is interpreted. Thus, subconscious heuristic processing of features influences how a person interprets what they are reading and seeing.

Many studies of heuristic processing focus on subjective qualitative accounts of how individuals evaluate information (e.g. Fogg, 2003; Metzger et al., 2010; Metzger & Flanagin, 2013). A limited number of empirical studies have examined the types of heuristics triggered during credibility evaluation. For example, Bellur and Sundar (2014) describe how a popular logo could excite heuristics related to credibility but that the logo itself is simply a contextual feature, not the cognitive shortcut likely to be triggered. Likewise, the level of expertise of a source may be considered a feature for example, the postnominal qualification of PhD. However, such a feature will not always lead to the elicitation of the expertise (authority) heuristic. Based on these findings, Bellur and Sundar (2014) conclude that the lack of direct measurement of heuristics is due to them being frequently conflated with the features of online content when in fact they are not the same.

Simply manipulating a feature of online content may not always trigger an associated heuristic (Sundar, 2008). Likewise, the same features can excite different heuristics in different situations (Metzger et al., 2010). On this basis, it is reasonable to suggest that exploitation of context and content features should increase the likelihood of related heuristics being excited, particularly if multiple cues can be presented simultaneously.

The applicability of a heuristic is determined by its perceived relevance to a given decision-making situation and its perceived reliability in light of previous application. Depending on the situation and related personal experience, some heuristics have greater perceived reliability than others (Bellur & Sundar, 2014). The ability of online content features to prompt a heuristic response is also determined by the perceived intensity of presentation and its impact, which may vary along a continuum as opposed to simply being present or absent. Bellur and Sundar (2014) suggest that detection of online features should not necessarily require use of nominal categories (i.e. present or absent) but may instead adopt a scale to capture the 'intensity of the feature', thus reflecting the likelihood of triggering the heuristics.

2.4 | Detection of dis- and misinformation

Technological developments have made some progress towards identifying and managing dis- and misinformation online. For instance, network analysis algorithms can detect activity clusters within inter-connected websites on topics of political and social interest. Analysis of content within the online information environment to date has used natural language processing and sentiment analysis carried out by machine algorithms (Zhang et al., 2018; Zhou & Zhang, 2008). In the latter, words or combinations of words can be classified in order to draw inferences about the attitudes and opinions of populations or groups of users.

Current machine-based algorithms reduce the amount of information that individuals tasked with website deception detection and information assurance, such as intelligence analysts, might need to review. However, these algorithms are as yet unable to detect the nuanced use of verbal and non-verbal (e.g. images or video) content (Marcellino et al., 2017), nor adequately analyse the cultural contexts of posts (Farkas & Neumayer, 2018); hence, human scrutiny and judgement is required.

2.5 | The research gap

There is currently a gap in understanding the ways in which creators of online content can exploit heuristics and biases as CVs among consumers of content through manipulation of contextual and informational features. Furthermore, to the authors' knowledge no theoretical framework exists to provide insights into where website contents may have been designed to exploit common CVs amongst audiences; a form of social cognitive hacking. The present research contributes towards filling this gap through development of a working theoretical framework that aligns key features of online content to likely exploitable CVs. In doing so, four perspectives are brought together: (i) the psychology of credibility evaluation; (ii) the role of cognitive heuristics and biases in human information processing; (iii) IV assessment; and (iv) website user interface design guidance. Although the framework has been developed primarily for analysis of websites and website contents, it has the potential to be adapted for use with social media platforms.

3 | WORKING THEORETICAL FRAMEWORK

The inductive development of the framework began with a review of the background literature for each perspective. This was conducted using two parallel strands of activity. The first explored how heuristics and biases may be understood as CVs. The second strand focused on understanding how contextual and informational features shape the way in which information is perceived and understood. A key finding to emerge from the second strand was that online content can be manipulated through the presentation and organisation of contextual features and informational content to influence recipients' judgement and subsequent actions (Shedroff, 2001). The literature offered various approaches to the critical evaluation of online information and its credibility (e.g. Fogg, 2003; Metzger et al., 2010; Metzger & Flanagin, 2013) however, these tended to be organised in single lists for examination of features. For this reason, they were not considered specific enough to support the development of a framework for identifying where social cognitive hacking may be occurring. Therefore, a new approach was required.

Online content features can trigger heuristic responses and biases (CVs) and the likelihood of a response being elicited is determined by the perceived intensity of the presentation of a feature and its impact (Bellur & Sundar, 2014). The intensity may vary along a continuum which in turn can form the basis of a scale for capturing the 'intensity of the feature', thus reflecting the likelihood of it triggering an heuristic. It may therefore be possible for website designers to manipulate contextual credibility features to increase their perceived intensity and their potential to trigger heuristics. If heuristics associated with contextual credibility features can be triggered, it becomes possible to trigger further heuristics associated with the information presented within any given context to influence judgements about its validity. Thus, contextual and informational features can work together to potentially bypass rational evaluation processes to affect decision-making and behaviour.

A novel approach and framework was conceived which has at its core two dimensions; CC and IV (see Figure 1). The framework is intended as a basis for the effective assessment of the perceived intensity with which CC and IV features are presented and the likelihood and potential impact of triggering associated CVs.

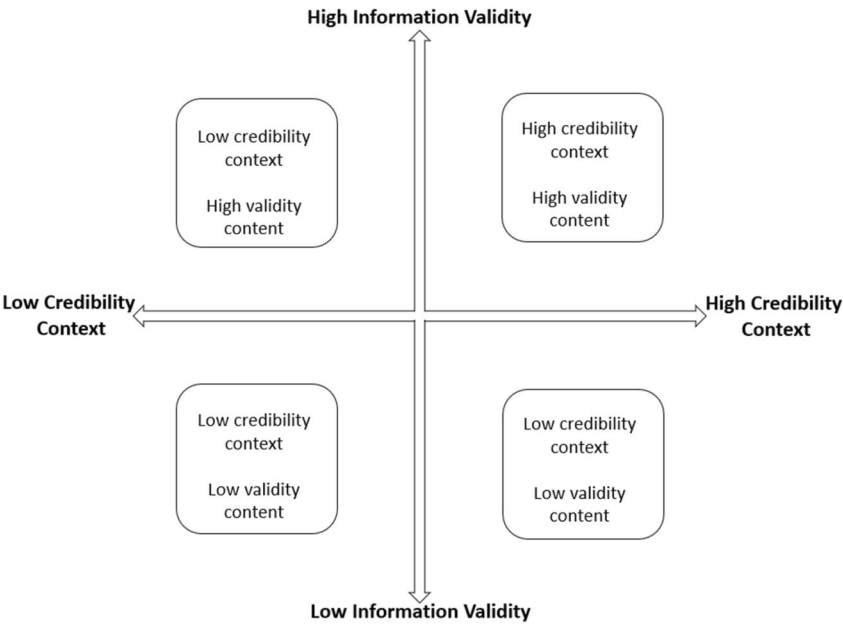


FIGURE 1 CC and IV dimensions. CC, context credibility; IV, information validity.

3.1 | Organisation of features along two dimensions

Candidate sets of CC and IV features were derived from the literature. The features in each set were selected on the basis of them having either a solid foundation in research into information credibility evaluation or from the literature on critical assessment of online content (e.g. Appelman & Sundar, 2016; Fogg, 2003; Harris, 2020; Metzger, 2007; Metzger et al., 2003; Metzger et al., 2010; Metzger & Flanagin, 2013; Warnick, 2004). In the context of this paper, features related to the credibility of the context (e.g. design quality, sponsorship, author/source information) were brought together and organised under the CC dimension. Features concerning the validity of content such as currency, objectivity, opinion, and verifiability were organised under the IV dimension.

Following an iterative process, the CC and IV features were refined and organised into categories on both dimensions. Explanatory descriptions for each CC and IV feature were generated to enable consistent analysis. For example, the category 'Trustworthiness' comprises numerous CC and IV features such as author/publisher information, language quality and verifiability. This process produced consolidated sets of 11 CC and 16 IV features. Within each dimension the features are considered to be mutually exclusive.

3.2 | Mapping of features

Since context frames the perception and interpretation of information the IV features were then mapped against the CC features. Mappings were assigned on the assumption that specific feature combinations were the more likely to influence information interpretation (see Figure 2).

The individual features on the CC and IV dimensions were then aligned with the heuristics and biases identified from the literature as CVs. The alignment of features to CVs was established inductively during a series of workshops to collectively analyse how different types of context and content, represented by the descriptions of the CC and IV features, potentially excite CVs. Table 1 presents the CVs aligned with the CC features and Table 2 presents the CVs aligned with the IV features. Note that the feature descriptions in Tables 1 and 2 are derived from

		CC Category											
		Purpose			Author / source verification mechanisms			Design & Assurance			Funding		
		CC Feature											
IV Category	IV Feature	Srl	1	2	3	4	5	6	7	8	9	10	11
Purpose	Clear purpose	1	x										
	Facts out of context	2	x										
	Clear audience	3	x										
Accuracy	Opinion	4					x			x		x	
	Rumour	5				x	x						
	Currency	6				x			x				
	Objectivity	7					x						
	Biased content	8								x		x	
Ambiguity	Memes	9									x		
	Satire	10	x										
Commercial	Commercial motive	11	x			x							
Dishonesty	Plagiarism	12		x			x						
Persuasion	Malicious persuasive intent	13					x	x			x		
Trustworthiness	Author/publisher information	14		x	x	x		x		x			x
	Verifiability	15		x	x	x		x		x			x
	Language quality	16		x	x	x	x	x		x			x

FIGURE 2 Matrix of propositional relationships between CC and IV features and their categories. CC, context credibility; IV, information validity; Srl, serial number.

TABLE 1 CC features and descriptions with associated CVs.

Srl	CC feature	Description	Associated CVs
1	Audience and purpose	Extent to which the intended audience can be inferred and extent to which the site's purpose can be inferred and evaluated in terms of its impact on credibility	None identified
2	Author/source information	Author/publisher identity and information content are clearly linked and the author/publisher is known and accepted as trustworthy with sufficient knowledge/expertise to write on the subject	Expectancy; reputation; trusted authority
3	Contact information	Presence of contact information for organisation/webmaster/authors/publishers etc	Expertise; trusted authority; reputation; organisational reputation
4	Anonymity	Presence of site mechanisms supporting publishing of anonymous content	Belief bias
5	Design quality	A well organised site has a logical layout that follows website design conventions	Information accessibility; expectancy
6	Sponsorship	Sponsorship information can be discerned. Sponsorship can influence the purpose of a website, the type of content presented, and the validity of the information. For this reason, it is useful to know who the sponsor is. Where the information cannot be discerned it may be necessary to use a reliable fact checking website	Trusted authority; endorsement (encompasses: Consensus; Bandwagon; conferred credibility; social proofing)
7	Automatic date stamp	Presence of date stamp showing date and time of content creation	Information accessibility; expectancy; availability
8	Governance information	Website features that include information about ownership (which differs from sponsorship) and policies and procedures for quality control and oversight	Trusted authority; organisational reputation; information accessibility; expectancy
9	Traceability of images used in site design	The provision of accurate information about the source and, where relevant, date of images increases the credibility of a website	None identified
10	Justifiability of advertising	The presence of advertising can reduce the credibility of a website unless the site provides justification for the presence of advertising. For example, the use of advertising may be balanced with organisation/author/publisher reputation, where production and hosting costs for valid information content have to be covered. Many mainstream news sites and high-quality blogs, for example, have advertising. This is often where the reader has not paid a subscription charge	Expectancy
11	Clickbait	'Clickbait' content can be identified by its use of language, for example, to stimulate curiosity or emotional engagement. It might also attempt to motivate the reader to take an action 'click	Affect

(Continues)

TABLE 1 (Continued)

Srl	CC feature	Description	Associated CVs
		here!' or use strategies such as 'headlines' that build suspense, sensation, lure and tease the reader into clicking on the content. The site owner receives revenue for every 'click'	

Abbreviations: CC, context credibility; CVs, cognitive vulnerabilities; Srl, serial number.

general website quality assessment criteria (e.g. Fogg, 2003; Harris, 2020) and occasionally academic literature such as (e.g. Metzger et al., 2010; Metzger & Flanagin, 2013). The CVs were derived exclusively from the academic literature (see Supporting Information S1: Appendix A for a full list with definitions and sources).

Each dimension could have been presented separately to assess the perceived intensity of the features thus reflecting the likelihood of their triggering associated CVs. However, because contextual features work with informational features to influence perception and interpretation, the two dimensions for assessing the perceived intensity of CC and IV features were brought together to form a quad-graph as an effective means for quantifying the combined effect of perceived feature intensity, CV triggering likelihood and potential impact.

To enable assessment of mapped pairs of features through analysing their location in any quadrant, two sets of questions and prompts for producing pairs of values along the scales for each dimension were required. Given the absence of empirical evidence regarding the extent to which different feature combinations trigger associated CVs, hypothesised weightings were developed and applied to initial score values from each question set such that the values on each dimension fall within the range −10 to 10. As such, the weightings were based on the authors' interpretation of the scientific and practitioner literature. Combinations of CC and IV feature weighted scores form pair values for quad-graph data-points to suggest the extent to which exploitation of CVs is occurring (see Figure 3).

Figure 3, with its example four data-points illustrates the quad-graph concept. Each of the four example data-points appears within a quadrant. Data-point A in Quadrant 1 (Q1) results from the indicative score values for a pair of features. The score of −6 for the CC feature, 'Governance information' (see Table 1, Srl 8) suggests low CC resulting from missing information about ownership and policies and procedures for quality control and oversight. (Note that ownership is not the same as sponsorship.) As can be seen in Table 1, the CVs aligned with this feature are: Trusted authority; Organisational reputation; Information accessibility; and Expectancy (see Supporting Information S1: Appendix A for definitions). The score of 6 for the IV feature, Author/publisher information (see Table 2, Srl 14) suggests that the information and the identity of the author or publisher are reasonably apparent and that the author/publisher is known and accepted as a trustworthy source with sufficient knowledge/expertise to write on the subject. The aligned CVs are: Cognitive authority; Trusted authority; Organisational reputation. The other three data-point examples can be interpreted in a similar way using the scores.

3.3 | Development of question sets and scoring system

The CC dimension has 11 features, and the IV dimension has 16 features. For each feature on each dimension there are one or more question items and prompts. These additional items and prompts are used to analyse elements of the features, which we refer to as 'feature elements'. These are used to assess either the presence, absence or the intensity of feature presentation acknowledging that the absence or low intensity of features can also influence CC and IV judgements (Bellur & Sundar, 2014). Responses are captured using a variety of formats. For some, a fixed nominal response is required where a categorical answer was deemed most appropriate to capture the presence or absence of a feature, for example, 'Yes', 'No', 'Sometimes' or 'N/A', etc. For example, CC Serial 7 (see Supporting

TABLE 2 IV features and descriptions with associated CVs.

Srl	IV feature	Description	Associated CVs
1	Clear purpose	Purpose of the information is clear that is, inform, argue, teach, sell, entertain, persuade etc	None identified
2	Facts out of context	Facts are not aligned with the apparent purpose	None identified
3	Clear audience	Extent to which the information/content is relevant to the intended audience and purpose of the website. [Researchers' assumption: Audience can be inferred from the purpose]	Availability; belief bias; salience; expectancy; perceived similarity
4	Opinion	The information presents a view or judgement not necessarily based on fact or knowledge (i.e. evidence) but based on for example, anecdotes/personal narratives/diatribe/rant. This is different from expert opinion	Perceived similarity of source; representativeness; cognitive authority
5	Rumour	Extent to which content contains rumour or urban myths/legends (stories or reports of uncertain or doubtful truth/unverified information/unverifiable information)	Novelty
6	Currency	The extent to which the currency of the information can be discerned	Availability
7	Objectivity	The content is consistent that is, free from contradiction, is factual, and the language used is balanced and unbiased. It is free from: Intemperate tone or language; over-claims; sweeping statements of excessive significance; conflict of interest	Expectancy; affect; salience
8	Biased content	Likelihood that the information presents a one-sided view that does not acknowledge opposing/alternative views or respond to them. The information appeals to a bias held by a person or a group accessing the information	Cross validation (encompasses consistency); cognitive authority; belief bias
9	Memes	The extent to which known internet memes are present. Memes are viral internet images containing short text, often perceived as humorous, where alternative connotations other than humorous intent can be identified. For example, memes may be used in an attempt to reinforce cultural beliefs by potentially activating alternative interpretations based upon cultural, political, racist beliefs amongst a subset of the wider audience	Empathy; affect
10	Satire	Extent to which satirical content masquerades as valid information can be discerned. Dis- and mis-information can masquerade as satire that is, hoax/inflammatory content with a motive other than for example, humour	Availability; truthing; belief bias; novelty
11	Commercial motive	Can commercial intent be discerned where the content (individual text/audio/video items, etc.) is selling or endorsing goods, services etc. for financial gain	Expectancy violation

(Continues)

TABLE 2 (Continued)

Srl	IV feature	Description	Associated CVs
12	Plagiarism	Information is plagiarised that is, copied from another source without attribution and presented as the author/publisher's own content	Belief bias
13	Malicious persuasive intent	The content seeks to influence decision-making and induce others to take a malicious course of action using emotional earnestness accompanied by exaggeration, sweeping generalisations, or absolutes. Content may include negative propaganda (e.g. anti-establishment) or pro-discrimination (e.g. culture, race, religion, disabilities) using in-group (e.g. us/we)/out-group (e.g. them) distinctions to dehumanise target out-groups and/or individuals' representative of an out-group	Cognitive authority; conformity; perceived similarity of source; identifiable victim effect; sense of urgency; empathy; affect; misconstrued altruism
14	Author/publisher information	Author/publisher identity and information content are clearly linked and the author/publisher is known and accepted as trustworthy with sufficient knowledge/expertise to write on the subject	Cognitive authority; trusted authority; organisational reputation
15	Verifiability	Content makes use of evidence from reputable sources which can be accessed either via links or reference information such that the information presented can be verified in other reliable source/s	Cross-validation (encompasses consistency)
16	Language quality	The appearance, professional quality of language used and clear writing style would score highly with an audience expecting information to be presented in this fashion. It adds to the transparency of the information. For example, high quality investigative journalism is different from low quality sensationalist journalism	Expectancy

Abbreviations: CVs, cognitive vulnerabilities; IV, information validity; Srl, serial number.

Information S2: Appendix B) asks, 'Does the technology used to create the site include an automatic date stamp showing time and/or date the content was added/posted? (Yes/No)'. Others require consideration of a list of check boxes next to feature elements to identify whether they are present or not. Where the purpose of the question is to elicit a judgement regarding the extent to which a feature impacts on either CC or IV judgements, that is, its 'intensity', responses are scored using subjective rating scales. Presence or absence of features may result in positively or negatively weighted scores. For example, the feature elements in IV Serial 7 (see Supporting Information S3: Appendix C) are intended to evaluate the objectivity of a website's content and their absence or presence will determine how they are scored. Further, presence or absence of a feature or feature element is occasionally used to initiate conditional branching within the CC and IV question sets.

Scales are either five-point (1–5), 7 point (1–7), 10-point (1–10), or 11-point (–5 to +5, with central option of '0'). The subjective rating scales for the features were developed by the authors. The type of scale used was decided on the basis of intelligibility of application to address specific feature elements. For example, an 11-point scale (–5

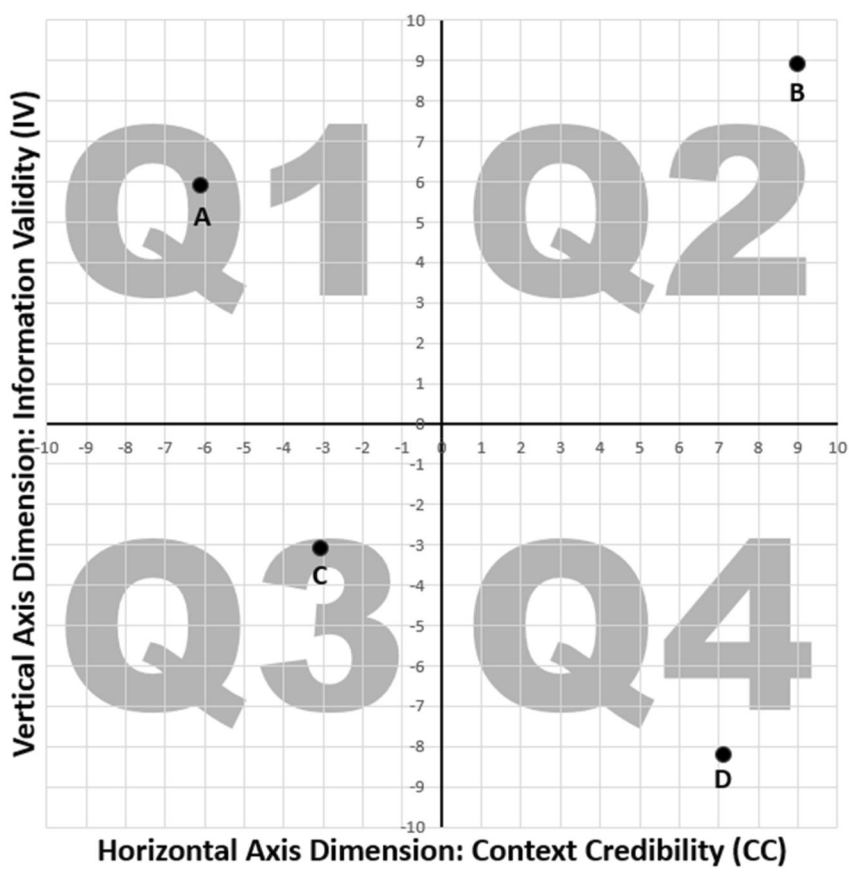


FIGURE 3 Indicative quad-graph with four data-points (A, B, C, and D) resulting from example CC-IV dimension value pairs located within quadrants (Q) 1–4 of the graph. CC-IV, context credibility-information validity.

to +5, with central option of ‘0’) was chosen for several feature elements because the use of positive and negative integers was considered more intelligible than descriptors such as ‘Medium positive impact’. A mid-point of ‘0’ was included to enable SMEs to indicate ‘No impact’ on judgement in their ratings. In contrast, a 10-point scale (1–10) was adopted for some feature elements. This allowed SMEs to rate the extent to which a feature impacts on either CC or IV judgements.

The value of responses to check box questions against a list of feature elements is TRUE or FALSE. These Boolean values are either converted into a numerical value or a text string to capture the feature elements that are present. Some of the questions require a free text response to generate summary information in support of the analysis; these were not scored. These questions represent an approach to indicate how additional descriptive information could be captured and used to populate a summary report for analysts. For example, CC Serial 1 captures the analyst’s inferences regarding the target audience and purpose of the website.

Responses to feature question items are captured and processed using formulae within a MS Excel spreadsheet to convert the initial numerical values into weighted scores. As previously described, weighted scores range between 10 and –10. The weighted scores from the CC and IV question sets produce 55 pairs of values based upon the mapping of CC-IV features in Figure 3. These value pairs are used to populate the quad-graph with data-points. (See Supporting Information S2: Appendix B and Supporting Information S3: Appendix C for screenshots showing the CC and IV question sets with prompts and response options as they appear within the MS Excel spreadsheet).

3.4 | Quadrant by quadrant implications for interpretation and judgement

Table 3 broadly defines the implications of feature combinations that appear in each of the quadrants as a result of conducting a structured website analysis. It should be noted that the feature combinations may be increasingly positive or increasingly negative thereby indicating higher or lower CC or IV. It is unlikely that all identified feature combinations would appear in a single quadrant. Therefore, overall judgement made about a website on the basis of the quad-graph would require further interpretation.

4 | PRIMA FACIE ASSESSMENT

A UK-based organisation provided a senior expert intelligence analyst as a subject matter expert (SME) for conducting a prima facie assessment (PFA) of the framework (including the question sets) to establish its subjective face validity. Professionally, the SME had responsibility for a team specialising in the identification and analysis of

TABLE 3 Implications of data-points resulting from positive (+ve) and negative (–ve) weighted values for CC-IV feature pairs being concentrated within Q, along with examples.

Q	CC value	IV value	Implication for analysis if data-points are concentrated within a quadrant	Example website types
Q1	–ve	+ve	Valid content is presented in a low credibility context. Few or no data-points in Q3 or Q4. The content on such sites is likely to be reliable even if it appears on a website with many negative context features (e.g. with advertising, low quality design, anonymous postings, etc.)	A scientist's personal research blog or an internet technology chat/online forum
Q2	+ve	+ve	Valid content is presented within a credible context. There may be data-points higher up in Q3 or Q4 but very few, if any, lower down. The content on websites with the majority of data-points concentrated within Q2 is likely to be reliable. However, the appearance of any data-points in any of the other quadrants, particularly Q4 suggests further scrutiny is required as the meaning of any verbal/nonverbal (images and video) content may be nuanced	Reputable non-aligned news portal or mainstream religious information site
Q3	–ve	–ve	Invalid content is presented in a low credibility. The majority of data-points appear in Q3, possibly with some to the left of Q4. Any data-points appearing in Q1 or Q2 are likely to be towards the bottom of the quadrants	Racially motivated website with discussion forum that includes abusive and/or emotive language, anonymous postings or website with sensational and/or 'clickbait' content
Q4	+ve	–ve	Invalid content is presented within a highly credible context. There is likely to be greater distribution of data-points within Q2 and Q4 and frequently with more in Q2 in order to give an overall plausible appearance. Analysis producing this type of patterns are likely to be passing off some invalid, deceptive/misleading information in a deliberate and targeted way; any verbal/nonverbal (images and video) content is likely nuanced	Ideologically driven news site or sites offering alternative clinically unproven medical treatments

Abbreviations: CC, context credibility; IV, information validity; Q, quadrants.

threats within website contents. (Ethical approval to proceed was granted through the researchers' university ethics system: ref. 9347/2019).

During the PFA, the SME (under the guidance of the research team), applied the framework to two websites that remain active at the time of writing. Website A is the English language version (it is available in multiple languages) of a European state-funded news and media network with global reach. Website B is a US based news and opinion website. The websites were proposed by the team as perceived vehicles for mis/disinformation, with Website A appearing to have far more subtle and nuanced content than Website B. This paper summarises the results from the PFA for Website A to demonstrate the output and reflect upon the steps taken to establish the framework's face validity. Note that the SME was familiar with Website A but had not carried out any formal analysis of it. The purpose of using a familiar website was to ascertain how closely the results of applying the framework matched the SME's expectations, including those formed previously, as a highly experienced website analyst.

Having applied the framework and question sets to Website A, the SME was asked to discuss their overall impressions of the framework, question sets and quad-graph and to comment on such aspects as its usability and utility. They were also asked to assess how well the output quad-graphs matched their expectations given their prior knowledge of Website A. The SME's conclusion was that it had high face validity and showed promise for integration within their working practices, if it were to undergo further development and validation. Furthermore, the SME confirmed that the distribution of data-points within the quad-graph and the implications in terms of how Website A presents and manipulates information met with their professional expectation. The SME stated that following further development the framework could provide analysts with a structured, standardised process for conducting formal analysis.

On completion of the PFA, the think-aloud protocols and additional SME feedback were analysed. The findings of the analysis were used to adjust usability issues and improve question set items to minimise ambiguity by making them more explicit. Where required, binary fixed nominal response formats were modified to ensure they were exhaustive of all possible states regarding the presence or absence of a feature. For example, the inclusion of 'sometimes' or 'not applicable (n/a)' as options. A sample of the original outputs from the SME's application of the method in the updated version of the proof-of-concept method spreadsheet is discussed in the following section.

4.1 | Website 'A' analysis

Figure 4 presents the output spreadsheet quad-graph for Website A. Note that only 44 of 55 possible data-points are visible because some have the same values as others and are therefore overlaid giving the appearance of being a single data-point.

Although there is reasonable distribution of data-points across Q1 and Q2 in Figure 4 consistent with a mainstream news media website (see Table 3 for implications), this discussion is limited to the analysis of Q4. The appearance of 24% of available data-points in this quadrant is considered unusual and suggests the presence of some form of cognitive hacking that is, a deliberate attempt to exploit CVs. Nearly a quarter of Website A's content appears to be of low or very low IV yet it is presented within a mostly credible context. With reference to the Q4 implications for analysis in Table 3, the distribution pattern suggests that the site may be subtly conveying invalid, deceptive/misleading informational content. The presentation of verbal (e.g. text or video with voice) and nonverbal (e.g. images and video without voice) content with low IV under the cover of a highly credible context is likely to be very nuanced and deliberately aimed at exploiting a set of CVs.

Table 4 lists pairs of weighted values for the CC and IV features in Q4. When used in conjunction with the features and CVs in Tables 1 and 2, it aids interpretation of the data-point distribution and indicates what CVs are possibly being targeted. The CC and IV features being used to do are shown. For example, we can see that the IV feature of 'malicious persuasive intent' (which is likely exploiting CVs such as cognitive authority,

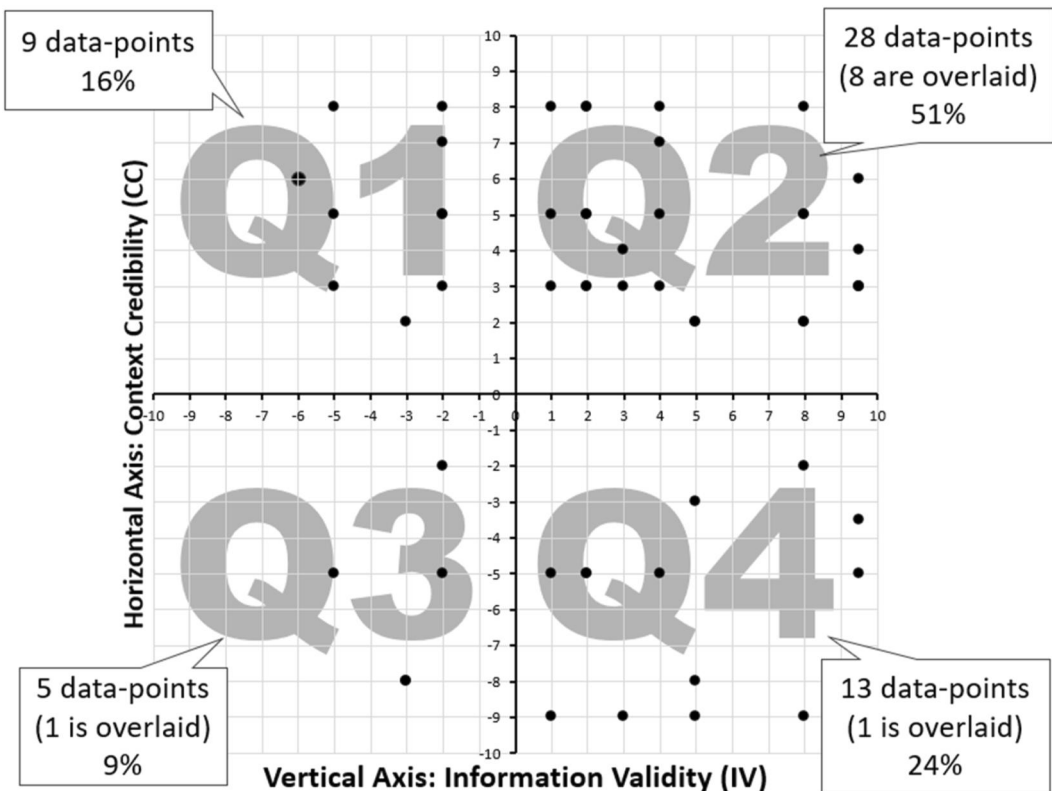


FIGURE 4 Output quad-graph showing distribution of data-points by (number and percentage) following application of the CC and IV question sets to Website A. CC, context credibility; IV, information validity.

conformity, and affect) has been detected within the website content with scores as low as -9 . Here, the associated CV of 'cognitive authority' is potentially targeted in a ploy to influence audiences to perceive low validity information as being trustworthy. It achieves this by suggesting that it comes from someone who is knowledgeable or informed in some way, even though this is unlikely to be true. If audiences believe that the information they have seen or read is trustworthy, it may subsequently influence aspects of their judgement, decision-making and behaviour. The other data-points with negatively scored IV features such as questionable author/publisher information, biased content, memes and misleading content masquerading as humour (IV feature 'Satire'), support an overall view of the website as one designed to exploit social cognitive processes (i.e. social CVs) to drive societal influence.

Regarding Q1 and Q2, the data-points in Q1 probably occur because CC and IV values indicate much of the website presents good quality information. However, there are some features on the website which attract a negative score such as anonymous contributions towards conversation threads. It should be noted that this is a common feature within many more balanced news media websites. For this reason, even highly reliable sites may produce a distribution of data-points across more than one quadrant upon application of the process. The data-points in Q2 indicate the presence of many CC-IV feature combinations (28 data-points, i.e. 51%) with positive weighted values, which initially creates a favourable impression of Website A. The subsequent large number of data-points in Q2 indicate the presence of valid information within a credible context. Thus, unwary audiences relying on System 1 heuristic thinking are less likely to detect the presence of the type of content that leads to data-points appearing in Q4. The 9% of data-points falling into Q3 could be interpreted as a deliberate appeal to audiences attracted by the types of website listed in the Table 3 examples.

TABLE 4 CC and IV data-point values and features from Website A, Q4.

CC, IV values/data-point	CC features	IV features
1, −9	Sponsorship	Malicious persuasive intent
1, −9	Sponsorship	Author/publisher information
3, −9	Audience and purpose	Malicious persuasive intent
5, −9	Traceability of images used in site design	Malicious persuasive intent
2, −5	Author/source information	Author/publisher information
2, −5	Governance information (quality control)	Author/publisher information
5, −8	Justifiability of advertising (clear)	Biased content (appeal to held bias)
4, −5	Contact information	Author/publisher information
8, −9	Design quality	Malicious persuasive intent
5, −3	Traceability of images used in site design	Memes
9.5, −5	Audience and purpose (credibility impact)	Commercial motive
9.5, −3.5	Audience and purpose (credibility impact)	Satire
8, −2	Design quality	Rumour

Abbreviations: CC, context credibility; IV, information validity.

5 | CONCLUSIONS

The novel theoretical framework described here makes an important contribution to the development of a structured, standardised process for conducting formal analysis. In some instances websites and their contents may be so nuanced that it is impossible for machine algorithms to detect that they may be subtly exploiting CVs. However, as shown by its ability to highlight data-points for Q4, the framework has promise in achieving this. It does so by indicating where CC features are positively reinforcing their associated CVs and combining them with negatively scored IV features (i.e. with low values).

Based upon the PFA, the framework is considered to have potential application in intelligence analysis. Beyond intelligence analysis, the framework’s organisation of features along the CC and IV dimensions will be of interest to teachers and students and others concerned with critically evaluating online information and how it may be organised to influence thought and behaviour. The framework could be used actively for designing more targeted and effective online marketing and communication campaigns. It will be of particular use in understanding how online content, that is at first impression visually and cognitively attractive, may be triggering CVs to intentionally bypass critical evaluation and reasoning. Finally, despite having been designed for human use, the underpinning theoretical framework has the potential to inform development of more sensitive machine-based algorithms for detecting dis- and mis-information.

5.1 | Limitations and future work

A major limitation of the current work is that due to availability, the PFA was conducted with a single analyst and therefore it was only able to establish the working framework’s face validity. Future work will build upon the research reported in this paper. First, it will refine the framework by establishing multiple SME judgements regarding the current hypothetical CC and IV feature mappings and associated weightings. Furthermore, it will indicate whether additional CC and IV features and combinations are required. Second, validation of the refined

framework will require several SME analysts to apply the framework to a set of nuanced websites identified by them during a series of workshops. The results from each analyst's application of the framework to the selected websites will be statistically analysed to determine their similarity. This will establish the wider reliability of the revised framework and identify where further amendments are required. Separate from determining reliability, experimental research could provide evidence for the probability that CVs associated with CC and IV feature combinations will be triggered.

ACKNOWLEDGEMENTS

This work was supported by the UK Defence Science and Technology Laboratory (DSTL) through the Defence and Security Accelerator (DASA) Phase 1 Behavioural Analytics Themed Competition. Contract number DSTLX1000133521.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

The data supporting this study cannot be made available due to commercial restrictions.

ORCID

Piers MacLean  <https://orcid.org/0000-0003-1307-9005>

Marie Cahillane  <https://orcid.org/0000-0001-8669-3407>

REFERENCES

- Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1), 59–79. <https://doi.org/10.1177/1077699015606057>
- Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 20:1–20:27. <https://doi.org/10.1145/3274289>
- Barfar, A. (2019). Cognitive and affective responses to political disinformation in Facebook. *Computers in Human Behavior*, 101, 173–179. <https://doi.org/10.1016/j.chb.2019.07.026>
- Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior*, 116, 106633. <https://doi.org/10.1016/j.chb.2020.106633>
- Bellur, S., & Sundar, S. S. (2014). How can we tell when a heuristic has been used? Design and analysis strategies for capturing the operation of heuristics. *Communication Methods and Measures*, 8(2), 116–137. <https://doi.org/10.1080/19312458.2014.903390>
- Cybenko, G., Giani, A., & Thompson, P. (2002). Cognitive hacking: A battle for the mind. *Computer*, 35(8), 50–56. <https://doi.org/10.1109/MC.2002.1023788>
- Dhoju, S., Rony, M. M. U., Kabir, M. A., & Hassan, N. (2019). Differences in health news from reliable and unreliable media. In *Companion proceedings of the 2019 world wide web conference* (pp. 981–987). <https://doi.org/10.1145/3308560.3316741>
- Doty, C. (2015). Social epistemology and cognitive authority in online comments about vaccine safety. In *Proceedings, the iConference 2015, Newport Beach Marriott Hotel & Spa, Newport Beach, California, USA, March 24-27, 2015*. University of Illinois. Retrieved from <http://hdl.handle.net/2142/73664>
- Ecker, U. K. H., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Pre-existing attitudes and continued influence of misinformation. *Memory & Cognition*, 42(2), 292–304. <https://doi.org/10.3758/s13421-013-0358-x>
- Farkas, J., & Neumayer, C. (2018). Disguised propaganda from digital to social media. In J. Hunsinger, L. Klasturp, & M. M. Allen (Eds.), *Second international handbook of internet research* (pp. 1–17). Springer. https://doi.org/10.1007/978-94-024-1202-4_33-1
- Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI '03 extended abstracts on human factors in computing systems* (pp. 722–723). <https://doi.org/10.1145/765891.765951>

- Friedman, H. H. (2017). Cognitive biases that interfere with critical thinking and scientific reasoning: A course module (SSRN Scholarly Paper ID 2958800). *Social Science Research Network*. <https://doi.org/10.2139/ssrn.2958800>
- Georgiadou, E., Rahanu, H., Siakas, K., McGuinness, C., Edwards, A., Hill, V., Khan, N., Kirby, P., Cavanagh, J., & Knezevic, R. (2018). Fake news and critical thinking in information evaluation. Western Balkan Information Literacy Conference WBILC 2018, 21-22 June 2018, Bihac, Bosnia and Herzegovina, Middlesex University.
- Grace, R., & Leskovich, R. (2015). Challenges in the use of social media data for the next generation analyst. *Next-Generation Analyst III*, 9499, 949903. <https://doi.org/10.1117/12.2177834>
- Harris, R. (2020). Evaluating internet research sources. Retrieved October 20, 2020, from <https://www.virtualsalt.com/evaluating-internet-research-sources/>
- Helmus, T. C., Bodine-Baron, E., Radin, A., Magnuson, M., Mendelsohn, J., Marcellino, W., Bega, A., & Winkelman, Z. (2018). *Russian social media influence: Understanding Russian propaganda in Eastern Europe*. RAND Corporation. Retrieved from https://www.rand.org/content/dam/rand/pubs/research_reports/RR2200/RR2237/RAND_RR2237.pdf
- House of Commons, Digital, Culture, Media and Sport Committee. (2019). Disinformation and 'fake news': Final report. Retrieved from <https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, 50(1), 46–70. <https://doi.org/10.1111/j.1460-2466.2000.tb02833.x>
- Lee, S. Y. (2018). Effects of relational characteristics of an answer on perceived credibility of informational posts on social networking sites: The case of Facebook. *Information Research*, 23(3), paper 796. Retrieved from <http://informationr.net/ir/23-3/paper796.html>
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, 66(1), 799–823. <https://doi.org/10.1146/annurev-psych-010213-115043>
- Marcellino, W., Smith, M. L., & Paul, C. (2017). *Monitoring social media. Lessons for future Department of Defense social media analysis in support of information operations*. RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR1742.html
- Martens, B., Aguiar, L., Gomez-Herrera, E., & Mueller-Langer, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3164170>
- Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the Association for Information Science and Technology*, 58(13), 2078–2091. <https://doi.org/10.1002/asi.20672>
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59(Part B), 210–220. <https://doi.org/10.1016/j.pragma.2013.07.012>
- Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. (2003). Credibility in the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. In P. Kalbfleisch (Ed.), *Communication yearbook 27* (pp. 293–335). Lawrence Erlbaum.
- Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439. <https://doi.org/10.1111/j.1460-2466.2010.01488.x>
- Niehoff, L. M., & Shah, D. (2017). The resilience of noxious doctrine: The 2016 election, the marketplace of ideas, and the obstinacy of bias. *Michigan Journal of Race & Law*, 22(2), 243–271. <https://repository.law.umich.edu/mjrl/vol22/iss2/3>
- Pantazi, M., Hale, S., & Klein, O. (2021). Social and cognitive aspects of the vulnerability to political misinformation. *Advances in Political Psychology*, 42(S1), 267–304. <https://doi.org/10.1111/pops.12797>
- Paul, C., & Matthews, M. (2016). *The Russian "firehose of falsehood" propaganda model*. RAND Corporation. Retrieved from https://www.rand.org/content/dam/rand/pubs/perspectives/PE100/PE198/RAND_PE198.pdf
- 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. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, 29(3), 343–373. https://doi.org/10.1207/s15516709cog0000_20
- Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science & Technology*, 41(1), 307–364. <https://doi.org/10.1002/aris.2007.1440410114>
- Samoilenko, S. (2018). Strategic deception in the age of "truthiness". In I. Chiluba (Ed.), *Deception and deceptive communication: Motivations, recognition techniques and behavioral control* (pp. 129–168). Nova Science Publishers.
- Shedroff, N. (2001). An overview of understanding. In R. S. Wurman (Ed.), *Information anxiety 2* (pp. 27–29). QUE.
- Soon, C., & Goh, S. (2018). Fake news, false information and more: Countering human biases. IPS Working Papers, 31. Retrieved from https://lkyspp.nus.edu.sg/docs/default-source/ips/ips-working-paper-31_fake-news-false-information-and-more_260918.pdf

- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger (Ed.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press. Retrieved from <https://doi.org/10.1162/dmal.9780262562324.073>
- Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news”. *Digital Journalism*, 6(2), 137–153. <https://doi.org/10.1080/21670811.2017.1360143>
- Tsang, S. J. (2022). Biased, not lazy: Assessing the effect of COVID-19 misinformation tactics on perceptions of inaccuracy and fakeness. *Online Media and Global Communication*, 1(3), 469–496. <https://doi.org/10.1515/omgc-2022-0037>
- Vidgen, B., Taylor, H., Pantazi, M., Anastasiou, Z., Inkster, B., & Margetts (2021). *Understanding vulnerability to online misinformation*. The Alan Turing Institute. Retrieved from https://www.turing.ac.uk/sites/default/files/2021-02/misinformation_report_final1_0.pdf
- Waltzman, R. (2017). *The weaponization of information: The need for cognitive security*. Testimony. RAND Corporation. Retrieved from https://www.rand.org/content/dam/rand/pubs/testimonies/CT400/CT473/RAND_CT473.pdf
- Warnick, B. (2004). Online ethos: Source credibility in an ‘authorless’ environment. *American Behavioral Scientist*, 48(2), 256–265. <https://doi.org/10.1177/0002764204267273>
- Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., Karger, D., & Mina, A. X. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion proceedings of the ‘the web conference’ 2018* (pp. 603–612). <https://doi.org/10.1145/3184558.3188731>
- Zhou, L., & Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM*, 51(9), 119–122. <https://doi.org/10.1145/1378727.1389972>

AUTHOR BIOGRAPHIES

Piers MacLean is a Lecturer in Human Cognition within Cranfield Defence & Security's Centre for Electronic Warfare, Information, and Cyber. Piers originally joined Cranfield University in 2007. Since 2010 his interest in the human behavioural sciences has resulted in extensive collaboration with Dr Marie Cahillane and together in 2013 they formed the Applied Psychology Group. During the past 13 years he has contributed to many Defence and Security funded research projects applying cognitive psychology in areas such as sequence acquisition in weapon handling, and debiasing techniques for intelligence analysts. More recent work includes Skill Fade (HSSRC 2019–2022); and Detecting the Exploitation of Human Cognitive Vulnerabilities Online (DASA Phase 1 and 2 Behavioural Analytics Competition, 2019–2021).

Dr Marie Cahillane (CPsychol) is Head of the Applied Psychology Group and a Senior Lecturer in Applied Cognitive Psychology within the Centre for Electronic Warfare, Information and Cyber at Cranfield University. She is a Chartered member of the British Psychological Society, with over 15 years' experience in leading and collaborating on defence and Security research. Defence and Security funded research as the principal investigator includes: Horizon scanning - enhancing individual and collective training (R-Cloud, 2023); Skill Fade (HSSRC 2019–2022); and Detecting the Exploitation of Human Cognitive Vulnerabilities Online (DASA Phase 1 and 2 Behavioural Analytics Competition, 2019–2021). Marie was also co-investigator on a recent project measuring sympathetic views towards extremist terrorist activities (UIC2, 2022).

Dr Victoria Smy (CPsychol) is a Lecturer in Applied Cognitive Psychology at Cranfield University with over 8 years of experience in Defence and Security (D&S) research. She specialises in the exploration of psychometrics (attitudes, beliefs, motivations, etc.), training (design, delivery, assessment, evaluation) and decision-support (tools, tactics and procedures). Victoria was principal investigator on a recent University Innovation Concept (UIC2, 2022) project, ‘Measuring sympathetic views towards extremist terrorist activities’. Other relevant Defence and Security research includes ‘Horizon scanning—enhancing individual and collective training’ (R-Cloud, 2023), ‘Human vulnerabilities in the face of cognitive hacking’ (DASA Phase 1 2019), ‘Skill fade’ (HSSRC 2019), and ‘Analytics Training—Cognitive Processes/Cognitive Biases in Decision Making’ (DHCSTC 2015).

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: MacLean, P., Cahillane, M., & Smy, V. (2024). Lifting the lid on manipulative website contents: A framework mapping contextual and informational feature combinations against associated social cognitive vulnerabilities. *Social and Personality Psychology Compass*, e12947. <https://doi.org/10.1111/spc3.12947>



DSPACE

<https://dspace.org/>

Lifting the lid on manipulative website contents: a framework mapping contextual and informational feature combinations against associated social cognitive vulnerabilities

MacLean, Piers; Cahillane, Marie; Smy, Victoria

2024-02-15

Wiley

<https://doi.org/10.1111/spc3.12947>

MacLean P, Cahillane M, Smy V., (2024) Lifting the lid on manipulative website contents: a framework mapping contextual and informational feature combinations against associated social cognitive vulnerabilities. *Social and Personality Psychology Compass*. Volume 18, Issue 2, February 2024, Article number e12947