ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI Computing

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0003

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

AI-Generated Misinformation and the Psychology of Belief Formation: Investigating Cognitive Biases, Emotional Manipulation, and Susceptibility to Digital Radicalization

Saba Nadeem1,* and Almas Irfan1

1 Department of Psychology, Rawalpindi Women University

Corresponding author: Saba Nadeem (sabahkhan1130@gmail.com) · ORCID: [0000-0000-0000-0000]

Author emails: Almas Irfan (almas.irfan@f.rwu.edu.pk)

ABSTRACT

Costs involved in the generation of fluent written material, imagery, and audiovisual manipulation have been reduced with the advent of Generative AI. This creates a potential risk for belief formation and diffusion. While cognitive biases, emotional manipulation, and online radicalization have all been studied individually, there is little knowledge about the relationship between exposure to content, belief formation, susceptibility to emotional manipulation, and radicalization through AI-generated material. This paper develops a theory-informed conceptual framework explaining the process through which cognitive biases and emotional manipulation affect perception of reality and belief formation, intentions to share, and attitude susceptibility towards radicalization. It is a conceptual framework paper supported by a preregistration protocol for a randomized online experiment. The theoretical framework includes dual-process theory, motivated reasoning, illusory truth effect, social identity theory, and emotional appraisal integrated into a moderated mediation model. The framework yields five testable hypotheses and treats exposure, belief, endorsement, and participation as different dependent variables.

Keywords: AI-generated misinformation; belief formation; cognitive biases; emotional manipulation; generative artificial intelligence; digital radicalization; attitudinal susceptibility; media literacy

Highlights

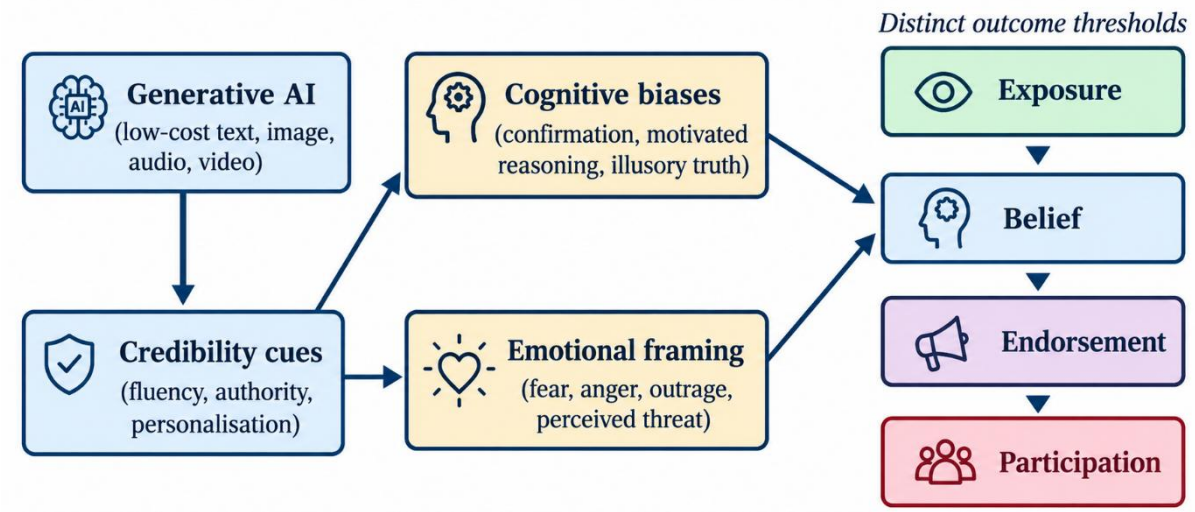
- Costs involved in the generation of fluent written material, imagery, and...
- This creates a potential risk for belief formation and diffusion.
- While cognitive biases, emotional manipulation, and online radicalization have...

IMPACT STATEMENT

Costs involved in the generation of fluent written material, imagery, and audiovisual manipulation have been reduced with the advent of Generative AI. This creates a potential risk for belief formation and diffusion. While cognitive biases, emotional manipulation, and online radicalization have all been studied individually, there is little knowledge about the relationship between exposure to content, belief formation, susceptibility to emotional manipulation, and radicalization through AI-generated material. This paper develops a theory-informed conceptual framework explaining the process through which cognitive biases and emotional manipulation affect perception of reality and belief formation, intentions to share, and attitude susceptibility towards radicalization. It is a conceptual framework paper supported by a preregistration protocol for a randomized online experiment. The theoretical framework includes dual-process theory, motivated reasoning, illusory truth effect, social identity theory, and emotional appraisal integrated into a moderated mediation model. The framework yields five testable hypotheses and treats exposure, belief, endorsement, and…

1 I. INTRODUCTION

Generative artificial intelligence technology enables the production of persuasive written text, imagery, audio, and video to be inexpensive and rapid [1]; yet, it presents a psychological problem that cannot be addressed through technical identification alone: in what conditions do people end up believing, spreading, and acting on this information? These issues will be addressed by this paper in the following five stages. First, the paper will define AI-created misinformation and differentiate it from disinformation. Second, it will describe the process of belief formation and updating. Third, it will identify a particular gap in the literature regarding cognitive bias, emotional manipulation, and vulnerability to radicalization. Fourth, it will outline the research objective and objectives.



1.1 **Figure 1. Overview of the problem space, from AI-generated content to distinct outcome thresholds.**

As shown in Figure 1, generative AI influences human beliefs and behavior through credibility cues, cognitive biases, and emotional framing, potentially leading to endorsement and participation. However, exposure to AI-generated content does not necessarily result in belief or active participation.

1.2 A. Background and Context

The abilities of generative AI systems to generate fluent text, photorealistic images, synthetic voices, and fake videos inexpensively, as well as customize messages according to the language, tone, and preferences of the target audience, make it easier to create misleading content and undermine the usefulness of traditional authenticity cues [2]. This paper will examine the psychological processes involved in the consumption of such content. Specifically, it will analyze how and to what extent biases and emotions influence the

evaluation of AI-generated misleading information. However, in this paper, we will not assume that AI-generated information is always misleading or that its consumption leads to beliefs and further actions.

1.3 B. Psychology of Belief Formation

The process of evaluating information involves both rapid intuitive and slow effortful processes. The cues of credibility include expertise, reputation, fluency, familiarity, social validation, and congruence with previous beliefs. In addition, they often do not undergo checks. Information that agrees with the beliefs of an individual and is delivered by a trusted in-group source is more likely to be accepted and analyzed less than other information. Repeating information makes it easier to process, which is why it may be taken for truth. Generative AI technologies can provide many of these cues inexpensively [5]; therefore, the process of reception of information will largely determine the impact of the technology.

1.4 C. Research Problem and Knowledge Gap

Earlier studies have considered three topics: (i) the cognitive psychology of accepting misinformation via the mechanisms of confirmation bias, motivated reasoning, and illusory truth effect; (ii) the role of emotions and emotional moral language in attracting attention and sharing; and (iii) social-psychological theories of radicalization. There are three gaps that require research. First, these three topics have not been merged into one framework that includes such outcomes as perceived truth, sharing, and susceptibility to radicalization. Second, the literature on AI-generated information mainly concerns detecting AI-generated content, while the interaction between synthetic credibility cues and cognitive/emotional factors requires more research [6]. Third, radicalization has mainly been viewed as an outcome of exposure to information, but not of endorsement. Emotional and cognitive biases have been known to influence the evaluation of misinformation; however, their interaction with regard to AI-generated content is unknown.

1.5 D. Research Objectives and Questions

The broad objective is to design an evidence-based theoretical model of the role that cognitive biases and emotional manipulation play in belief in AI-generated misinformation and susceptibility to digital radicalization. More specific objectives are: (O1) to estimate the effect of AI-generated misinformation on the perception of truth and belief; (O2) to estimate the effect of confirmation bias, motivated reasoning, and familiarity on the perception of credibility; (O3) to estimate the effect of emotional framing on processing and sharing intention; (O4) to assess the relationship between these factors and susceptibility to radicalizing narratives; and (O5) to design a model and test it. The following are four research questions to be answered in this study. RQ1: How strongly do cues of credibility and repetition predict perceptions of truth and beliefs? RQ2: Is there a moderating influence of belief congruence (which is used as a proxy for confirmation bias and motivated reasoning) on the predictive relationship? RQ3: What is the role of emotional framing in arousal, perception of credibility, and intention to share? RQ4: What is the role of arousal and identity-threatening situations in radicalization susceptibility besides beliefs?

1.6 E. Research Contributions

Contribution of the study: The contribution of the current paper is that it provides an empirically testable model that takes into account the combination of cognitive and emotional mechanisms in terms of their impact on three different types of outcomes (belief, sharing, and susceptibility). Additionally, the current paper provides recommendations on how to measure exposure and susceptibility, and not confuse one for the other.

2 II. LITERATURE REVIEW

However, the review will be more about synthesis and comparison of relevant literature rather than analysis of individual works. After presenting the basic types of AI-based misinformation and their perceptions of credibility and detectability, the review will move to discussing cognitive biases, framing, and social identity in terms of how they contribute to the acceptance of belief, attention to and forwarding of messages, and digital radicalization. In the course of discussion, established facts will be distinguished from hypotheses, and contradicting views will be identified.

2.1 A. AI-Generated Misinformation

Different forms of AI-generated misinformation exist. Generated synthesized texts may look credible because of fluency and authoritative style, and in one study, GPT-3-generated disinformation was evaluated to be more persuasive than human-written disinformation. Fake evidence (documents, citations, or statistics) may rely on credibility heuristics and anchoring; explainability-oriented language-model frameworks strive to increase the auditable nature of generated claims [7]. Deep fakes and manipulation of audio-video material depend on the credibility of recorded evidence and are especially believable in this sense. As for personalized narratives, they adjust to the interest and language of the in-group (what is technically achieved through user-profiling and tracking techniques [8]), and are expected to trigger confirmation and in-group biases, although the superiority over generic texts should not be taken for granted. The generation technologies are well-researched [9]; however, their psychological effects remain largely unknown. The human ability to detect generated material is weak: in one study, participants were unable to reliably differentiate between tweets written by GPT-3 and by real users.

2.2 B. Cognitive Biases and Belief Acceptance

Table 1 is devoted to mechanisms rather than synonyms. Confirmation bias deals with selective search and interpretation of evidence; motivated reasoning is goal-driven and identity-protective processing; the illusory truth effect is a fluency-based consequence of repetition that does not require any direction; availability and anchoring bias judgment because of easy recall and initial anchor points; authority and in-group biases act via source cues. Literature points to the strength of the repetition effect and congruence with existing beliefs, but differs with regard to the influence of analysis on identity-protective cognition and the length of time that changes in belief last.

Table 1. Cognitive biases and belief formation

Bias	Psychological meaning	Role with AI-generated content (illustrative example)	Boundary condition
Confirmation bias	Selective seeking and interpretation of evidence that supports existing views	A fluent AI-written article echoing a user's view is accepted without checking its claims.	Weaker when accuracy goals are salient or disconfirming evidence is easy to obtain
Motivated reasoning	Goal-directed processing that protects identity or preferred conclusions	Dismissing an accurate correction because it threatens group identity	Stronger for identity-relevant topics; weaker for low-salience issues
Illusory truth effect	Repeated statements feel truer because they are processed more fluently	An AI-generated claim recirculated across accounts feels familiar and therefore credible	Attenuated when prior knowledge is strong, or source is known to be unreliable
Availability heuristic	Judging frequency or risk by ease of recall	Vivid synthetic images of an event inflate perceived prevalence	Weaker when base-rate information is salient
Authority bias	Deference to perceived expertise	Fabricated expert quotation or professional formatting	Weaker when source is verifiably absent or flagged

Bias	Psychological meaning	Role with AI-generated content (illustrative example)	Boundary condition
In-group bias	Preferential trust in members of one's own group	Message adopting in-group idiom and symbols	Stronger under identity threat; weaker with cross-group norms of scrutiny

As can be seen from Table I, several psychological biases play an important role in perception and reaction to content created by AI. Confirmation bias, motivated reasoning, the illusory truth effect, availability heuristic, authority bias, and in-group bias could affect credibility perceptions and belief reinforcement. Yet, depending on different conditions (e.g., previous knowledge, personal relevance of the topic, trust in the source, access to counterevidence), their impact may vary.

2.3 C. Emotionally Manipulative Techniques

Fear, anger, anxiety, threat appeals, humiliation, and perceived injustice could attract attention, make people resort to heuristic cues, and stimulate sharing. In particular, moral-emotional language has been found to be associated with increased diffusion of political messages on social media, having higher impact in ideologically homogeneous networks than heterogeneous ones. Emotional persuasion is not a manipulative technique itself – manipulation is defined as a use of emotions for undermining reasoning or achieving deception. Emotions themselves may convey truthful information, and sharing, belief formation, and attitude change are not necessarily interconnected phenomena (people may share information without believing it). The question of whether AI-created emotional manipulation will be more efficient than human messaging is open, and the majority of the existing literature on the issue discusses technical attacks [10].

2.4 D. Social Identity and Digital Radicalization

In this case, digital radicalization is defined as a process of growing extremism of beliefs, attitudes, and behavior that promotes intergroup conflict at least partially through digital means [11]. Radicalization online does not necessarily mean exposure to radical content or endorsement of extremism and extremist acts, which are considered distinct phases of the process. Social identity, narratives of grievances, injustice perception, isolation, sense of belonging, and community endorsement can be considered potential influential factors. Group identity and threat theory, in particular, are especially significant among influential theories; yet the proportion of each factor, including exposure to algorithms, self-selection, and offline risk, still needs to be clarified [12]. There is no group specifically vulnerable to radicalization, and most people exposed to radical content are not radicalized.

2.5 E. Critical Synthesis and Research Gap

Previous established results: repetition and consistency with previous beliefs affect truth perception, and emotions affect attention and sharing. Previous contested topics: the general role of analytical thinking, the extent of identity protective effects, and the relationship between sharing and belief. Issues regarding research design: single-session studies, convenience sampling, laboratory artifacts, self-reports, and immediate results. Understudied topics: combined effect of credibility heuristics, repetition, and emotional frames on AI-produced information, and the effect of AI literacy. Prior AI-threat research has also focused more on cybersecurity and national defense [13], [14] rather than on individual beliefs. Inferences are hard when only mediators are measured, and susceptibility is difficult to measure due to low frequency and unethical experimental manipulation of radicalization, so attitude indicators are problematic.

2.6 F. Selected Theories

Dual process theory describes two types of processing modes – intuitive and analytic. It explains the intuitive endorsement of fluency and authority, but it is criticized for being too simple a model and for failing to identify when analytic effort will occur. Both theories of motivated reasoning and confirmation bias posit a directional bias in processing, while directional bias may arise from prior-weighted updating as well. Hence, the concept of belief congruence serves as a moderating variable. The illusory truth effect predicts that repetition leads to perceived truth because of fluency, but its impact depends on the degree of prior knowledge. Social identity theory posits that group membership influences trust and attitudes towards other

groups, relating identity threat to the readiness to believe grievance narratives; however, it fails to explain why a very small percentage of people turn into extremists. The theories of emotional appraisal argue that threat, injustice, and loss lead to the emotional reaction that drives behavior and see motivation as an intermediate variable; however, manipulation and persuasion cannot be distinguished between them.

2.7 G. Proposed Conceptual Model

In order to bring together the psychological theories described above, the following conceptual model is presented as an attempt to explain the effect of generative AI on human behavior through cognitive biases, emotional framing, and perception of credibility [15]. This model separates exposure, belief, endorsement, and participation as distinct results of AI impact and does not treat exposure as something that will inevitably lead to the behavioral effect. In addition, it takes into account such boundary conditions as previous knowledge, identity relevance, reliability of the source, and goals of accuracy.

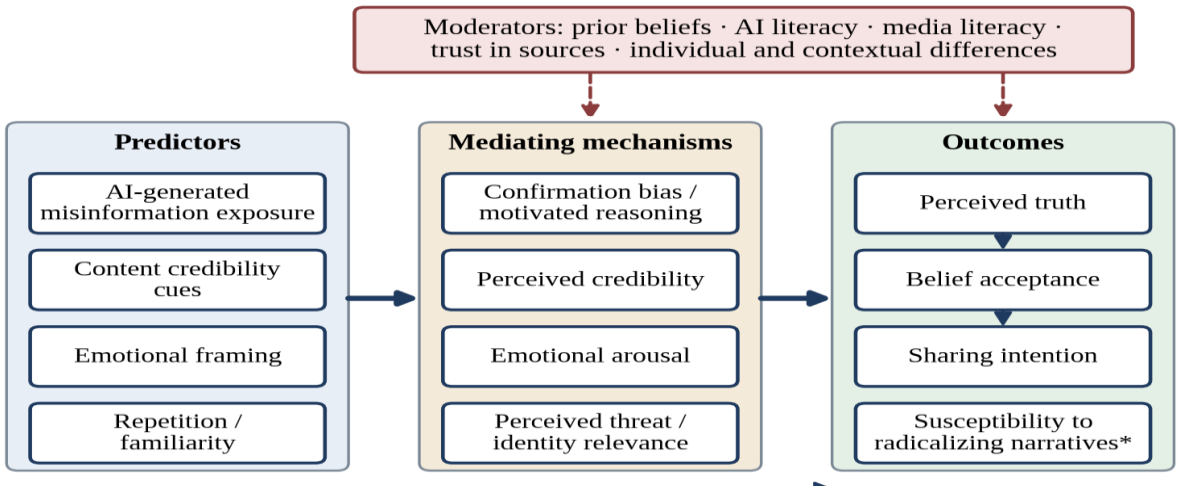


Figure 2. Conceptual framework of AI-generated misinformation and belief formation (hypothesized relationships).

The model is shown in Figure 2 below. Predictors (left side): Misinformation generated through AI technology, credibility cues, framing, repetition. Mediators (middle): Confirmation bias/motivated reasoning, perceived credibility, emotional arousal, threat/identity relevance. Outcome variables (right side): Perceived truthfulness, belief acceptance, sharing intentions, susceptibility to polarizing stories [16]. The solid lines represent the predicted relationships, whereas the dashed lines represent the predicted conditional relationships from the moderators (beliefs, media literacy, source credibility, situational factors). This figure captures the predicted relationships that will be empirically tested in the experiments as part of the hypothesis testing paths.

2.8 H. Research Hypotheses

The following research hypotheses are based on the proposed conceptual framework and theoretical concepts of cognitive psychology, morality, and AI-human interactions. These hypotheses explore the impact of AI-generated misinformation, perceived credibility, cognitive biases, and emotional framing on the process of belief formation, endorsement, and engagement [17, 18]. The hypotheses also take into account the moderating effect of individual and situational factors. Table 2 summarizes the five hypotheses along with justifications, variables, and analysis.

Table 2. Proposed hypotheses and evaluation methods

Hypothesis	Theoretical justification	Variables involved	Proposed analytical method
H1. Credibility cues increase belief acceptance, partly via perceived credibility.	Dual-process theory; authority and credibility heuristics	Credibility cues → perceived credibility → belief acceptance	Mediation with bootstrapped indirect effect (Eq. 2–3)
H2. Belief congruence	Motivated reasoning;	Exposure × belief	Moderation regression

Hypothesis	Theoretical justification	Variables involved	Proposed analytical method
strengthens the effect of exposure on perceived truth.	confirmation bias	congruence → perceived truth	with simple-slopes (Eq. 1)
H3. Emotional framing increases arousal, and arousal predicts sharing intention above and beyond belief.	Emotional appraisal; moral-emotional communication	Framing → arousal → sharing; belief as covariate	Mediation; hierarchical regression
H4. Repeated exposure increases perceived truth, less so for high-AI-literacy or high-prior-knowledge participants.	Illusory truth effect; fluency	Repetition × literacy / prior knowledge → perceived truth	Mixed-effects model; moderation
H5. Arousal and identity-relevant threat are positively associated with attitudinal susceptibility beyond belief acceptance.	Social identity theory; emotional appraisal	Arousal, threat, belief → susceptibility	Hierarchical regression; SEM if justified (association only)

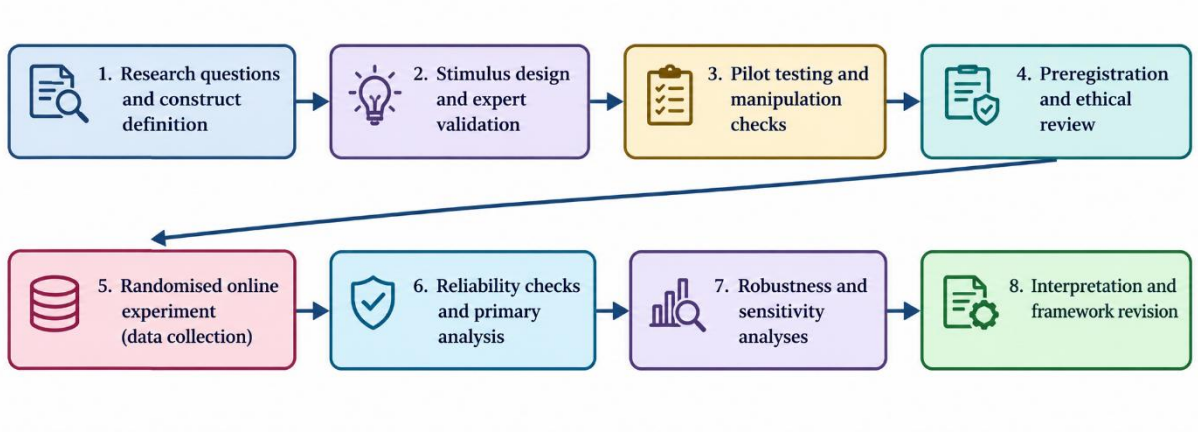
As shown in Table 2, the five research hypotheses explore the connections between the cues of credibility, congruence of beliefs, emotional framing, repetition, and susceptibility of individuals to AI-generated material. Each research hypothesis has a theoretical foundation in psychological theories and presents an appropriate way of analysis through mediation analysis, moderation regression, mixed-effects modelling, and structural equation modelling. The combination of these five research hypotheses provides the basis for an empirical examination of the suggested conceptual model.

3 III. METHODOLOGY

This section presents a reasonable and ethical approach to test the suggested framework. This includes a justification of the methodology and explanations of what each element is capable of showing: experimentally manipulated factors and relations between variables that are measured. In addition, the population, sampling, procedure, constructs, measurement, analyses, along with generic formulas for moderation and mediation will be outlined. Measures of validity, reliability, and safety for the participants will also be provided. No data have been collected yet.

3.1 A. Research Design

Two designs are used: conceptual framework development (Sections I–II), which formulates propositions through theoretical synthesis, and a proposed protocol for testing them, a preregistered randomized online experiment (not yet conducted). The experiment will manipulate emotional framing (neutral vs. high-arousal) and credibility cues (absent vs. present) between subjects, and repetition (single vs. repeated exposure) and message version (AI-generated vs. human-written) within subjects. Manipulated factors support experimental inference; measured mediators, moderators, and susceptibility support associations only (see Figure 3).



3.2 **Figure 3. Proposed research workflow from construct definition to interpretation.**

As illustrated in Figure 2, the research methodology follows an eight-step process to ensure systematic investigation, methodological rigor, and reliable findings. The process begins with defining research questions and constructs, followed by stimulus design, expert validation, pilot testing, and preregistration with ethical review. Subsequently, a randomized online experiment is conducted for data collection, followed by reliability assessment, primary analysis, and robustness checks. Finally, the findings are interpreted to refine the proposed conceptual framework and support the development of evidence-based conclusions.

3.3 B. Population, Sampling, and Procedure

Target population: adults (18+) who regularly use social media and are fluent in the language of the study; participants who fail the preregistered attention check are excluded. Recruitment will use online panels with quotas. The final sample size has not been set; it must be established before data collection by an a priori power analysis for the smallest effect size of interest (e.g., G*Power for ANOVA and regression; Monte Carlo simulation for mediation; α = .05; power .80–.90; allowance for attrition). Procedure: consent; baseline measurement of prior beliefs and moderators (separate from stimulus exposure); random assignment; exposure to brief, matched, social-media-like messages on low-stakes disputed issues; measurement of outcomes and manipulation checks; and debriefing with corrections. Stimuli (AI-generated or human-written, by condition) will be checked by experts for equal length, readability, and factual status, and pretested for emotional intensity. No actual extremist recruitment material will be employed; the radicalization narratives will only be mild and fictionalized grievance-and-blame narratives.

3.4 C. Variables, Definitions, and Measures

The constructs are defined in Table 3 and distinguished between theoretical constructs and their operationalization's. For the variables for which there is no validated scale (AI literacy, perceived credibility, narrative susceptibility), the scale would need to be selected and adapted/validated prior to conducting the experiment (pre-test, factor analysis, validation). The self-report approach is vulnerable to the issue of social desirability and acquiescence; solutions to minimize these biases include anonymity, reverse keying, and attention checks. Susceptibility would be measured using attitudes towards the grievances and an appropriate measure of radicalism intention, yet to be selected.

Table 3. Research constructs and operationalization

Construct	Conceptual definition	Operational definition and measurement	Analytic role
Exposure; credibility cues; framing; repetition	Features of AI-generated content and its presentation	Manipulated conditions (AI vs. human version; cue present/absent; neutral vs. high-arousal; 1 vs. 3 exposures); manipulation checks	Independent variables
Confirmation bias /	Selective acceptance of belief-congruent	Belief congruence (baseline attitude × message direction); congruence-	Moderator/mediato

Construct	Conceptual definition	Operational definition and measurement	Analytic role
motivated reasoning	information	weighted evaluation index	r
Perceived credibility	Judged trustworthiness and accuracy of message and source	Multi-item rating scale (to be validated); reliability by α or ω	Mediator
Emotional arousal; threat/identity relevance	Activation and perceived danger to self or group	Self-reported arousal and discrete emotions; threat and identification items	Mediators
AI literacy; media literacy; source trust	Knowledge and dispositions that condition evaluation	Literacy tests (AI literacy requires development); trust scale	Moderators
Perceived truth; belief acceptance; sharing intention	Judged veracity; agreement held with confidence; stated intention to share	Truth rating; agreement × confidence; sharing-intention items (not behavior)	Outcomes
Susceptibility to radicalizing narratives	Attitudinal receptivity to grievance-based, us-versus-them messaging	Receptivity scale; validated intention instrument (to be selected). Not exposure or behavior	Outcome (association only)

As shown in Table 3 below, the research operationalizes the major constructs that are necessary in the investigation of how the AI-generated content impacts the judgment and behavioral intentions of human beings. In this case, the independent variables, mediators, moderators, and dependent variables have been identified and defined, along with a proposed method of measurement for each of them, to ensure a rigorous scientific investigation. These constructs include perceptions about the credibility of the message, emotional arousal, belief acceptance, intentions to share, and susceptibility to the radicalizing narrative.

3.5 D. Analytical Methods

Data analysis begins with descriptive statistics, testing for missing data using multiple imputation if data are missing randomly, conducting reliability tests (alpha or omega), and conducting manipulation checks. The effects of manipulations are measured using ANOVA or mixed models. The generic model for moderation analysis is

$$Yi = \beta 0 + \beta 1Xi + \beta 2Wi + \beta 3(Xi \times Wi) + \beta 4Ci + \varepsilon i \quad (1)$$

where Yi is the dependent variable of participant i, Xi is the predictor (e.g., exposure), Wi is the moderator (e.g., prior-belief congruence), Ci is the set of covariates, β is the coefficient vector to be estimated, and εi is the error term. Mediation is modeled as follows:

$$Mi = \alpha 0 + aXi + \alpha 2Ci + eMi \quad (2)$$

$$Yi = \gamma 0 + c'Xi + bMi + \gamma 2Ci + eYi \quad (3)$$

where Mi is the mediator, a is the effect of X on M, b is the effect of M on Y controlling for X, c′ is the direct effect, and a×b is the indirect effect, which will be assessed with bootstrapped confidence intervals. These are generic models to be specified in terms of the selected variables; no coefficients have been estimated. Because M is measured rather than manipulated, a×b can be interpreted causally only under additional assumptions (no unmeasured M–Y confounding, including confounders affected by X; correct temporal order; correct model specification); temporal ordering alone is not sufficient. Even where X is randomized, indirect effects will be reported as consistent with, not proof of, mediation. Structural equation modeling will

be employed only if measurement models fit well. Effect sizes with confidence intervals, Holm correction of p-values across confirmatory tests, and sensitivity analyses (alternative scoring, exclusion rules, and covariates) will be reported.

3.6 E. Reliability, Validity, and Ethics

Construct validity is supported by theory-driven operationalization, review of the stimuli by experts, and convergent and discriminant validation. Internal validity depends on random assignment and manipulation checks. External validity is constrained by the artificial nature of exposure and the sampling frame. Reliability is based on α or ω; selection bias is diagnosed through attrition analysis; reproducibility is supported by preregistration and open data/materials. Institutional review is mandatory before data collection (not yet obtained); consent includes the right to withdraw; data are minimized and anonymized; extremist recruitment content is avoided; participants receive signposted support; and debriefing refutes any false claims. Reporting is aggregate and non-stigmatizing.

4 IV. RESULTS

This section presents a conceptual synthesis and a proposed framework for evaluating the five hypotheses. Because the experiment has not been conducted, no participant data, statistical results, or simulations are reported, and nothing here should be read as an empirical finding. The section (i) traces the proposed pathway from exposure to belief updating, (ii) states, for each hypothesis, evaluation criteria that would support, qualify, or contradict it once data exist, (iii) discusses theoretical contributions and practical implications, and (iv) states limitations and future research.

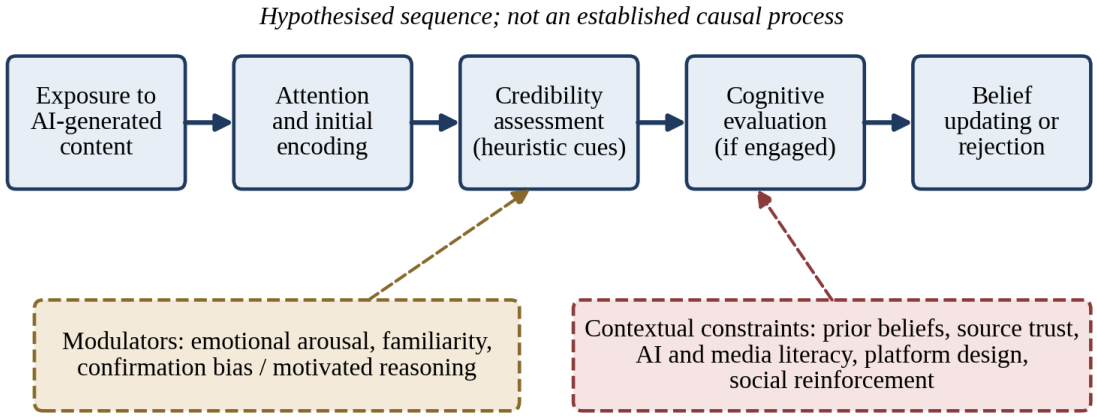


Figure 4. Proposed psychological pathway from exposure to belief updating (theoretical, not an established causal sequence).

The suggested process is depicted in Figure 4. Stages of this process will be affected by such factors as arousal, familiarity, and confirmation bias. Prior beliefs, trust towards the source, literacy, medium design, and social context will be the conditions for the stages. In this framework, exposure, perception of truth, belief acceptance, endorsement, sharing intention, susceptibility to radicalizing narratives, and radicalization itself or participation are regarded as different outcomes. Exposure does not imply perception of truth, sharing does not require belief, and susceptibility is not necessarily radicalization. It is believed that cognitive biases, arousal, credibility, repetition, prior beliefs, and social identity can interact with each other in the following way: repetition and credibility increase fluency and perceived credibility; prior beliefs and identity determine acceptance or rejection of such cues; and arousal, especially in case of identity threat, decreases scrutiny and increases sharing and acceptance of grievances. These relations are proposed and are not verified.

The criteria used for evaluation of hypotheses are provided in Figure 5. H1 (credibility cues ? perceived credibility ? belief acceptance) is supported by credible cue effect on belief and reliable indirect pathway through perceived credibility (it is conditioned by the measure of credibility), contradicting it in case of absence of this effect; H2 (exposure × belief congruence ? perceived truth) is supported by interaction when exposure increases the level of perceived truth in relation to congruent material, qualifying or contradicting it in absence of interaction or its restriction to identity-related content; H3 (framing ? arousal ? sharing, controlling for belief) is supported by a direct effect of arousal on sharing intention above belief acceptance,

qualifying it if both effects exist, contradicting it if the effect is absent; sharing intention is not sharing behavior. H4 (repetition × literacy or prior knowledge? perceived truth) is supported by the effect of repetition that is weakened with a higher level of literacy/knowledge, contradicting it in case of a constant effect; lack of validation of the literacy measure is a limitation. H5 (arousal and identity-relevant threat? attitudinal susceptibility beyond belief) is supported by association remaining after control for belief acceptance, contradicting it if it disappears; presence of association does not indicate causality.

Hypothesis	Would support	Would qualify	Would contradict
H1	Cues raise belief via credibility	Cue effect with no credibility mediation	No cue effect on belief
H2	Reliable exposure × congruence interaction	Interaction only for identity-relevant topics	No interaction (accuracy motive favoured)
H3	Arousal predicts sharing beyond belief	Arousal predicts both sharing and belief	Arousal unrelated to sharing
H4	Repetition raises truth; attenuated by literacy	Repetition effect equal across literacy levels	No repetition effect
H5	Arousal and threat linked to susceptibility	Link vanishes after controlling for belief	No association (not causal in any case)

Figure 5. Conceptual evaluation map (not empirical results) of result patterns that would support, qualify, or contradict each hypothesis.

Figure 5 is conceptual and contains no data. It substitutes possible patterns that would support, modify, or refute each hypothesis in lieu of illustrative interaction graphs that could be confused with empirical findings to ensure the framework remains falsifiable.

4.1 A. Possible Susceptibility Mechanisms and Alternative Explanations

It may be that an increased level of arousal resulting from the identity threat would make one more susceptible to believing the narrative of grievances. H5 will test only an association within one session. While an observed association between arousal, identity threat, and attitudinal susceptibility would not prove that misinformation leads to radicalization, attitudinal susceptibility is neither radicalization nor engagement in it. Possible alternative explanations include pre-existing beliefs, demand effects, overlap of measures, and omitted variable bias. Absence of difference between AI-generated and human-written messages in terms of susceptibility would indicate the importance of psychological processes, not the origin of information.

4.2 B. Interpretation and Theoretical Contributions

The process of acceptance of misinformation by AI is not governed by any one mechanism. Heuristic fluency, repetition, and authoritative signals work largely on a heuristic level, while congruence and identity influence the evaluative goals; emotions channel attention and motivation towards sharing, which is independent of belief formation. By bringing together both cognitive and emotional processes and making distinctions between outcome levels, the model perceives generative AI as a provider of credibility cues rather than an agent of psychology. This model also admits the possibility that congruence is a result of prior weight updating through rational means, thereby negating any possibility of a theory solely based on motivated reasoning. In the context of radicalization, this model differentiates between susceptibility and radicalization.

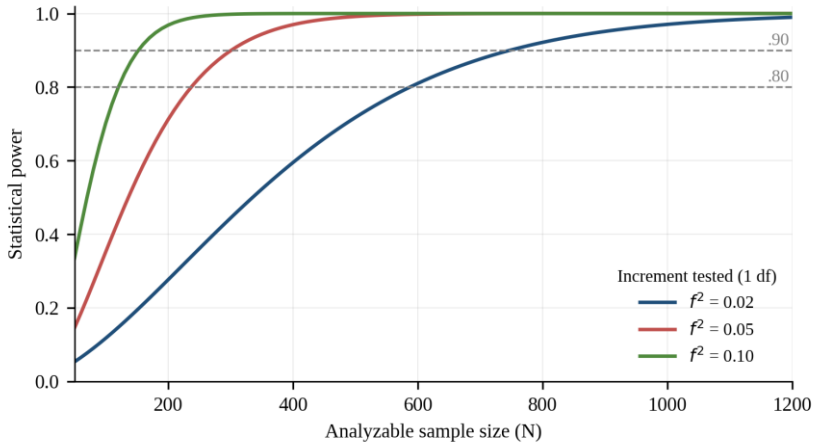


Figure 6. Power of a single tested regression term as a function of analyzable sample size (α = .01; six predictors; f^2 = .02, .05, .10). Design-planning calculation, not empirical results.

Figure 6 shows how strongly the required sample depends on the assumed effect. For f^2 = .10, power reaches .80 at about 121 participants and .90 at about 153, whereas for f^2 = .02 it remains below .80 until roughly 588 participants. The planning assumption about the smallest effect of interest therefore determines the scale of the study more than any other design choice.

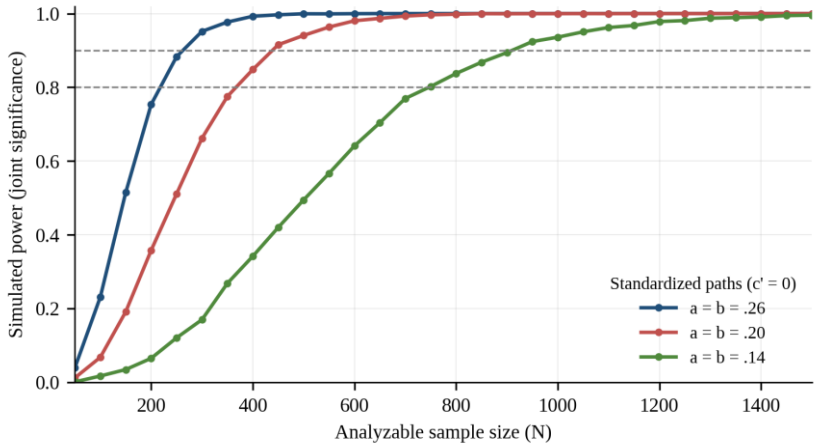


Figure 7. Simulated power of the joint-significance test of an indirect effect (α = .01 per path; 4,000 replications per point; standardized paths a = b; direct effect set to zero). Design-planning simulation, not empirical results.

Figure 7 shows that mediation requires considerably more participants than a single-term test of comparable standardized size. The implied indirect effects (a × b) are .0196, .04, and .0676 for paths of .14, .20, and .26; simulated power reaches .80 at about 750, 370, and 220 participants, respectively. Because the mediators are measured rather than manipulated, adequate power would not remove the causal-interpretation limits noted in Table 4.

Table IV. Sensitivity analysis: smallest effects detectable at selected analyzable sample sizes

Total N	d (α = .05, power .80)	d (α = .01, power .80)	d (α = .01, power .90)	f^2 (1 df, α = .01, power .80)
200	0.40	0.49	0.55	0.059
300	0.32	0.40	0.45	0.039
400	0.28	0.34	0.39	0.029
600	0.23	0.28	0.32	0.020
800	0.20	0.24	0.27	0.015

As shown in Table 4, the statistical power analysis illustrates how increasing the sample size improves the ability to detect smaller effect sizes under different significance levels and power requirements. The results indicate that a sample of 200 participants may be sufficient for detecting moderate effects, whereas larger samples of 400–800 participants enable the detection of smaller effects, particularly under stricter significance thresholds. These estimates provide preliminary guidance for determining an appropriate sample size for the proposed study. However, the final sample size should be established through a formal power analysis that considers the primary hypothesis, experimental design, statistical test, multiple comparisons, and expected participant attrition.

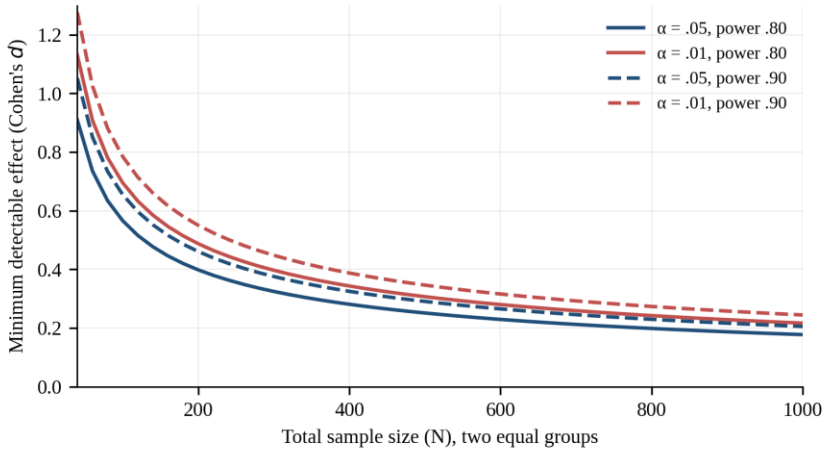


Figure 8. Minimum detectable effect (Cohen's d) for a two-group comparison as a function of total sample size, by significance level and power. Design-planning calculation, not empirical results.

Table 6 and Figure 8 invert the question and ask what the design could detect. With 400 analyzable participants, a between-subjects factor comparison can detect d ≈ 0.34 at α = .01 and power .80, and a single regression term can detect f^2 ≈ 0.029. Effects smaller than about d = .30 would therefore not be reliably detected below roughly 400–600 participants, and a null result in a smaller sample would not be evidence of no effect.

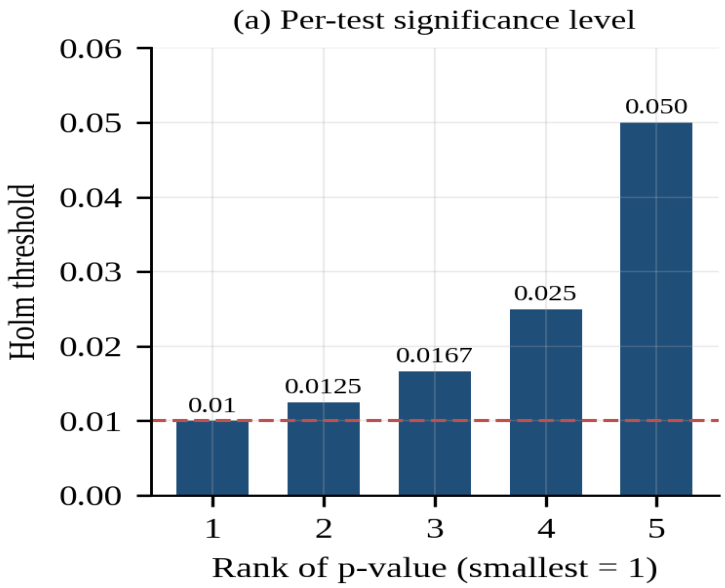


Figure 9. Holm step-down thresholds for the five confirmatory tests (a) and the analyzable sample required at each threshold for $f^2 = .05$ and power

As shown in Figure 9, the figure illustrates the Holm–Bonferroni correction thresholds for five ranked p-values, arranged from the smallest to the largest. The significance threshold increases progressively from 0.010 to 0.050, while the red dashed line represents the conventional significance level of 0.01. This correction controls the family-wise error rate when multiple hypotheses are tested, reducing the likelihood of false-positive findings. Each p-value should be compared with its corresponding adjusted threshold to determine statistical significance, with the sequential procedure stopping when a comparison fails.

.

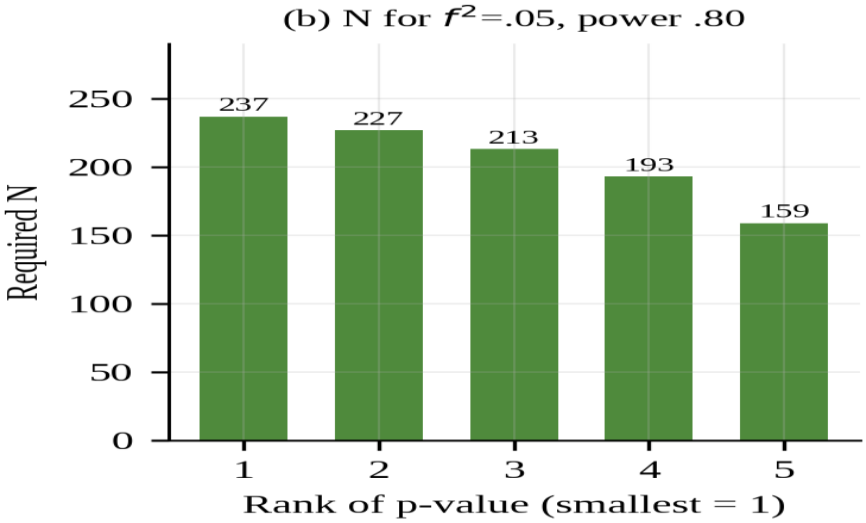


Figure 10. Required sample size (\(N\)) for detecting an effect size of \(f^2 = 0.05\) at 80% statistical power across five ranked hypothesis tests.

Figure 10 shows the required sample size to detect an effect of \(f^2 = 0.05\) at a statistical power of 0.80 on five hypothesis tests ranked according to their p-values. The required sample size ranges from 237 participants for the p-value of the lowest-ranked test to 159 participants for the fifth-ranked p-value. The presented example provides preliminary guidance on sample-size calculation; the required sample size should always be calculated in the framework of a power analysis corresponding to the statistical model used for the main study analysis.

4.3 C. Practical Implications

If future empirical work will support the presented framework, it can provide a theoretical foundation for such practices in artificial intelligence and media literacy that would address heuristic use (such as lateral reading and accuracy prompts), design responsible AI to incorporate content provenance and disclose the synthetic nature of AI-produced information, ethical moderation of emotional content as a signal of potential review and not falsehood, and preventive digital education. There is no silver bullet against disinformation and radicalization, and each measure needs to be independently evaluated. The effectiveness of the framework was not demonstrated.

4.4 D. Limitations and Future Research

The main limitation of the work is evidentiary: the experiment has not been done yet, so the suggested links need further empirical confirmation, including the determination of the strength of these links and for which individuals they work. Other limitations include the challenge to detect AI-produced content in the wild, the difference between perceived and actual credibility, measurement of the belief shift, potential self-reporting bias, the dependence on the specific topic and platform, lack of cross-cultural generalizability, and susceptibility vs. actual radicalization gap. Future research should include the preregistered experiment, valid measurement of constructs, longitudinal and field studies, cross-cultural replication, and intervention evaluation.

5 V. DISCUSSION

This suggested theoretical model depicts the influence of generative AI on human decision-making and intention to act through credibility cues, cognitive biases, and emotional framing. The biases of confirmation, motivated reasoning, and repeated exposure could be the elements causing the perceptions of truthfulness and belief, while the activation of emotions and threats to identity could result in intentions to endorse and spread AI-based content. Still, these influences will be affected by the existing prior knowledge, AI literacy level, source trustworthiness, and individual motives. The distinctions among the constructs of exposure, belief, endorsement, and participation create an opportunity for a more detailed understanding of the reaction to the generated information. In addition, appropriate sample design, validity of used measurement tools, multiple test adjustment, and robustness checks are needed in this case. As the proposed hypotheses should be confirmed in further study, it is essential to empirically test their relationships conceptually.

6 VI. CONCLUSION

In this paper, we argue that generative AI could increase the impact of cognitive biases and emotions in the assessment of information credibility, as well as reduce the costs of generating manipulative stories, though not everyone is expected to be equally affected by this technology. The paper makes a conceptual contribution since the vulnerability to digital radicalization is not determined by exposure to information alone but involves multiple factors and should be analyzed at least through five distinct stages, including exposure, perception of the truth, belief acceptance, intention to share, and attitudinal susceptibility. The proposed framework combines both cognitive and emotional mechanisms, offers five research hypotheses, and describes an experiment design that is ethical. The relationships described in this framework are only theoretical propositions rather than proven claims.

REFERENCES

[1] A. Raza, M. Sana, T. Hanif, A. Ali, S. G. Virk, and M. S. Iqbal, “IntelliWriter: A multi-model AI framework for intelligent SEO-driven content generation and blog automation,” in *Proc. 11th Int. Conf. Next Generation Computing (ICNGC)*, 2025, pp. 61–64.

[2] A. López-Borrull and C. Lopezosa, “Mapping the impact of generative AI on disinformation: Insights from a scoping review,” *Publications*, vol. 13, no. 3, Art. no. 33, 2025, doi: 10.3390/publications13030033.

[3] U. K. H. Ecker et al., "The psychological drivers of misinformation belief and its resistance to correction," *Nature Reviews Psychology*, vol. 1, pp. 13–29, 2022, doi: 10.1038/s44159-021-00006-y.

[4] V. Vellani, S. Zheng, D. Ercelik, and T. Sharot, "The illusory truth effect leads to the spread of misinformation," *Cognition*, vol. 236, Art. no. 105421, 2023, doi: 10.1016/j.cognition.2023.105421.

[5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018, doi: 10.1126/science.aap9559.

[6] A. Ali et al., "The threat of deepfake technology to trusted identity management," in *Proc. 2022 Int. Conf. Cyber Resilience (ICCR)*, Dubai, UAE, 2022, pp. 1–5, doi: 10.1109/ICCR56254.2022.9995978.

[7] J. Udry and S. J. Barber, "The illusory truth effect: A review of how repetition increases belief in misinformation," *Current Opinion in Psychology*, vol. 56, Art. no. 101736, 2024, doi: 10.1016/j.copsyc.2023.101736.

[8] J. J. Van Bavel, S. Rathje, M. Vlasceanu, and C. Pretus, "Updating the identity-based model of belief: From false belief to the spread of misinformation," *Current Opinion in Psychology*, vol. 56, Art. no. 101787, 2024, doi: 10.1016/j.copsyc.2023.101787.

[9] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological Science in the Public Interest*, vol. 13, no. 3, pp. 106–131, 2012, doi: 10.1177/1529100612451018.

[10] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, vol. 188, pp. 39–50, 2019, doi: 10.1016/j.cognition.2018.06.011.

[11] K. F. Khan, A. Ali, Z. F. Khan, and H. Siddiqua, "Artificial intelligence and criminal culpability," in *Proc. 2021 Int. Conf. Innovative Computing (ICIC)*, 2021, pp. 1–7.

[12] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017, doi: 10.1257/jep.31.2.211.

[13] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science Advances*, vol. 5, no. 1, Art. no. eaau4586, 2019, doi: 10.1126/sciadv.aau4586.

[14] L. K. Fazio, R. M. Pillai, and D. Patel, "The effects of repetition on belief in naturalistic settings," *Journal of Experimental Psychology: General*, vol. 151, no. 10, pp. 2604–2613, 2022, doi: 10.1037/xge0001211.

[15] M. Thaler, "The fake news effect: Experimentally identifying motivated reasoning using trust in news," *American Economic Journal: Microeconomics*, vol. 16, no. 2, pp. 1–38, 2024, doi: 10.1257/mic.20220146.

[16] J. Li, "Not all skepticism is 'healthy' skepticism: Theorizing accuracy- and identity-motivated skepticism toward social media misinformation," *New Media & Society*, vol. 27, no. 1, pp. 522–544, 2025, doi: 10.1177/14614448231179941.

[17] R. A. Naz, S. Nadeem, A. Ali, H. Riaz, I. Tahir, and M. A. Khan, "Psychological manipulation in cyberspace: Social engineering as human-centric cyber threat," in *Proc. 11th Int. Conf. Next Generation Computing (ICNGC)*, 2025, pp. 345–348.

[18] A. Ali, K. F. Khan Ghouri, H. Naseem, T. R. Soomro, W. Mansoor, and A. M. Momani, "Battle of deep fakes: Artificial intelligence set to become a major threat to the individual and national security," in *Proc. 2022 Int. Conf. Cyber Resilience (ICCR)*, Dubai, UAE, 2022, pp. 1–5, doi: 10.1109/ICCR56254.2022.9995821.

Declarations

Funding. [Funder name and grant number, or a statement that no external funding was received.]

Conflicts of interest. [Relevant interests, or: The authors declare no conflicts of interest.]

Ethics and consent. [Approval reference / not applicable statement.]

Data availability. [Repository / access conditions, or not applicable.]

Code availability. [Repository / version, or not applicable.]

Author contributions. [CRediT roles and initials.]

Acknowledgments. [Optional. Delete if none.]

Use of generative AI. [Disclose AI tool use, or state that none was used.]

How to Cite This Article

SUGGESTED CITATION

S. Nadeem and A. Irfan, "AI-Generated Misinformation and the Psychology of Belief Formation: Investigating Cognitive Biases, Emotional Manipulation, and Susceptibility to Digital Radicalization," International Journal of Security, Intelligence, Governance & Modern AI Computing (IJSIGMA), vol. 1, no. 1, pp. 50–67, 2026, doi: [DOI to be assigned].